ORBIT: DUAL-SHELL LATENT GEOMETRY FOR DIFFERENTIALLY PRIVATE ANOMALY REPRESENTATION SHARING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharing learned representations is increasingly important in cross-institutional anomaly detection, where differential privacy (DP) protects the released embeddings. Under extreme class imbalance, full-dimensional latent DP can erase the rare-class structure that makes the shared representation useful. We show that no per-record ϵ-DP release that protects the class can exceed precision $e^\epsilon/(e^\epsilon + r)$ at imbalance ratio r, and that a two-point release of a single coordinate attains this bound. We propose ORBIT (One-dimensional Radial Budgeting with Inter-shell Topology), which concentrates class-separating structure into a bounded radial coordinate rather than perturbing the full latent representation. Its encoder is trained with a dual-shell margin objective that places normal and anomalous records around two target radii. ORBIT clips the radius to a fixed range and releases it through a one-dimensional Laplace mechanism whose sensitivity depends on that range rather than the latent dimension. Under the same per-record guarantee, full-dimensional latent DP stays near the base rate up to ϵ=20 on 577:1 fraud data, whereas ORBIT's released radius reaches AUPRC 0.74 at ϵ=10 (bound 0.97), on par with a supervised score released in the same way. On the tabular datasets, unsupervised one-dimensional summaries reach at most 0.43 at ϵ=10, so compression alone does not explain the gain. Unlike a supervised score, ORBIT's representation also carries a direction, which a scoped variant publishes. This variant lies outside the bound because it declares the class public, and its formal guarantee covers only the radius. The identity leakage of the direction is bounded under an angular-concentration condition and audited with direction-only attacks, which are consistent with chance on both tabular datasets but detect training-set membership of minority records on CIFAR-10. Under this scope, ORBIT's latent-space synthetic release retains the non-private AUPRC at ϵ=1 on all three datasets.

1 INTRODUCTION

Cross-institutional anomaly detection, such as fraud, intrusion, and abuse monitoring, increasingly relies on exchanging learned representations, or data synthesized from them, rather than raw records (Yang et al., 2019; Arachchige et al., 2020; Weggenmann et al., 2022; Perez et al., 2024). The released material must support downstream learning without access to raw records, so we measure utility by what a classifier learns from it: directly from released representations, and through train-on-synthetic-test-on-real (TSTR) (Esteban et al., 2017) for synthetic data. We target the supervised setting, in which anomaly labels or trusted indicators are available for shaping; the guarantee applies to the released representations, not to encoder training.

Under extreme class imbalance, a common protection, isotropic DP noise on every latent coordinate, can erase the rare-class structure that makes the shared representation useful. On 577:1 fraud data, for example, full-dimensional latent DP calibrated to a standard per-record guarantee keeps the rare-class AUPRC near the base rate (0.004, against a base rate of 0.002 and 0.86 without privacy) even at ϵ=20 (Figure 1). DP is known to harm minorities disproportionately (Bagdasaryan et al., 2019; Feldman, 2020), and synthetic data does not remove this effect (Stadler et al., 2022). It remains open why the degradation is so severe at the representation level and how privacy can be allocated without destroying the useful structure.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Two considerations determine what can be achieved. First, any per-record release that protects the class faces a limit: it satisfies $\mathrm{TPR} \leq e^{\epsilon}\,\mathrm{FPR}$, so its precision cannot exceed $e^{\epsilon}/(e^{\epsilon}+r)$ at imbalance ratio r (Proposition 1), and at 577:1 precision above $1/2$ requires $\epsilon > 6.4$. Second, isotropic perturbation spreads the budget over every coordinate, although the rare-class decision can be carried by far fewer coordinates. Compression alone is not the answer: under the same guarantee, a class-informative coordinate reaches AUPRC 0.74 at ϵ=10 on the same data, whereas unsupervised one-dimensional summaries reach at most 0.43. The released coordinate must carry the rare-class separation.

Selective protection is an established idea, as in Pufferfish/Blowfish (Kifer & Machanavajjhala, 2014; He et al., 2014), one-sided DP (Kotsogiannis et al., 2020), and label DP (Chaudhuri & Hsu, 2011; Ghazi et al., 2021). These approaches take the data structure as given and decide what may remain public. They do not address how a *learned representation* should be structured when the mechanism is meant to protect a particular coordinate. Metric-learned representations offer a starting point: angular margins concentrate the directions of each class (Deng et al., 2019), and the feature norm carries graded information such as sample quality (Meng et al., 2021) and out-of-distribution signal (Park et al., 2023). This suggests giving the two coordinates different functional roles, with the radius carrying the protected quantity and the direction released under a declared scope. The two are not disentangled, since both can carry class information, so whether the direction can remain public is evaluated separately.

This paper asks how a learned representation should be structured so that a per-record private release comes close to this limit, and what can be released beyond it under a declared scope. ORBIT (**O**ne-dimensional **R**adial **B**udgeting with **I**nter-shell **T**opology) trains the encoder with a dual-shell margin objective that structures normal and anomalous records around two target radii. It clips the radius to a fixed range and releases it through a one-dimensional Laplace mechanism, which satisfies standard per-record ϵ-DP and approximates the two-point release that attains the limit. A scoped variant also publishes the direction. The angular concentration that results from shaping provides the premise for a bound on the identity leakage of the direction, and we audit that leakage with direction-only attacks.

Contributions. (1) *Representation–mechanism co-design:* the dual-shell objective concentrates the protected, class-separating information in one bounded coordinate, the radius, while the direction keeps the remaining structure for other release modes. The inter-shell margin controls noise-induced class crossover, and the radial release window determines the sensitivity, which does not depend on the latent dimension. (2) *The imbalance limit as an evaluation yardstick:* we make explicit that per-record ϵ-DP releases that protect the class cannot exceed precision $e^{\epsilon}/(e^{\epsilon}+r)$, a direct consequence of the DP definition. A two-point release of one coordinate attains this bound, and a radial release with window width Δh and margin M needs a privacy budget larger by the factor $\Delta h/M$ to reach the same ratio of true to false positive rates. (3) *Evidence that the gain needs a task-structured coordinate:* under the same per-record guarantee, ORBIT's released radius reaches 58–76% of the limit on 577:1 fraud data (AUPRC 0.12 at ϵ=5 against 0.20; 0.74 at ϵ=10 against 0.97) and follows the limit across imbalance ratios, while full-dimensional latent DP stays near the base rate up to ϵ=20. Unsupervised one-dimensional summaries do not reproduce the gain, a supervised score matches it, and across the evaluated margins the radius utility grows with the margin, as the crossover analysis predicts. (4) *Representation-level release and a scoped audit:* under the direction-public scope, the released representation retains the non-private AUPRC in latent-space synthetic release at ϵ=1. Publishing the direction, a declared scope outside the limit, carries most of the scoped release's utility, so we audit it: direction-only attacks map an operating regime (both tabular datasets) and a failure regime (a 432-image minority on CIFAR-10).

2 Preliminaries and Threat Model

Instance-level DP. ORBIT enforces *instance-level* (per-record) DP on each latent representation before release; the guarantee is defined over pairs of possible individual representations rather than adjacent datasets. *Definition 1.* A randomized mechanism $\mathcal{M}$ satisfies instance-level ϵ-DP if for any two representations $z, z' \in \mathcal{Z}$ and any $S \subseteq \mathrm{Range}(\mathcal{M})$, $\Pr[\mathcal{M}(z) \in S] \leq e^{\epsilon}\Pr[\mathcal{M}(z') \in S]$. *Definition 2.* For a map h on the latent space, the instance-level global sensitivity is $\Delta h = \max_{z,z'} \|h(z) - h(z')\|$ over the domain of individual records. The encoder E_θ is proprietary and

never released; the guarantee covers the released representation z_{priv} only, in the single-release setting (repeated releases compose in the standard way (Dwork & Roth, 2014; Mironov, 2017)).

Threat model and the protected quantity. We assume a strong adversary with unbounded auxiliary information: it observes every released representation, knows the public mechanism parameters (shell radii, noise scale), and holds the complete records of all non-target individuals. Consistent with one-sided DP (Kotsogiannis et al., 2020), whatever the public direction carries, including the class, is treated as background information available to the adversary. Extreme imbalance concentrates the risk on the rare class, whose sparse, isolated representations may be particularly vulnerable to membership inference (Shokri et al., 2017); this motivates a privacy model that states *which quantity, if leaked, causes privacy harm*, rather than spending the budget uniformly. Following label DP and one-sided DP, we consider a selective setting in which coarse task-relevant information is authorized for release and finer-grained individual information is protected. What the receiving institution legitimately consumes is the class (normal or anomalous), which the scope treats as public. We designate *how atypical* a record is within its class as the graded, individual-level privacy-sensitive quantity. Feature norms can carry such graded information, for example sample quality and out-of-distribution signal (Meng et al., 2021; Park et al., 2023). On both tabular datasets, an anomaly's radial position within its shell correlates positively with a separately computed atypicality score (Appendix H). We therefore protect the radial coordinate carrying this designated privacy-sensitive quantity and release the angular coordinate as a declared public component whose safety is not assumed but bounded (Section 5) and audited (Section 6.4). The two coordinates play different functional roles but are not disentangled: both carry class information (Appendix H).

The protected coordinate. Every released representation is decomposed into a *public angular coordinate* $\vec{u} = \hat{z}/\|\hat{z}\|_2$ and a *protected radial quantity* $s = h(\hat{z})$, where $\hat{z} = \Pi(z)$ is the feasible-domain projection of Section 4.2; the mechanism publishes $\vec{u}$ unchanged and releases only a randomized version of s. *Definition 3 (Direction-public instance-level ϵ-DP).* A mechanism $\mathcal{M}$ satisfies direction-public instance-level ϵ-DP if, for every fixed public direction $\vec{u}$, the conditional radial mechanism $\mathcal{M}_s(s \mid \vec{u})$ satisfies instance-level ϵ-DP with respect to s: for any feasible s, s' sharing the same direction $\vec{u}$ and any $S \subseteq \mathbb{R}$,

$$\Pr[\mathcal{M}_s(s \mid \vec{u}) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}_s(s' \mid \vec{u}) \in S]. \tag{1}$$

This is a *scoped* guarantee in the tradition of Pufferfish/Blowfish, one-sided DP, label DP, and metric DP (Chatzikokolakis et al., 2013): we declare the scope first and enforce the standard guarantee over that scope. Two boundaries of this scope matter. First, publishing $\vec{u}$ is a policy choice, not a claim that it is harmless: Proposition 3 and Lemma 1 bound the identity information the direction can carry under an angular-concentration condition, an empirical risk that Section 6.4 audits directly. Second, Lemma 1 concerns identity and membership *conditional on the class label*; the class is deliberately public, and any residual within-class attribute signal falls outside the formal guarantee and is measured separately. The scope protects s, which includes the (public) class margin induced by the inter-shell separation as well as each record's radial position within its shell. It does not assert that the public coordinate is universally non-informative. Withholding $\vec{u}$ and releasing only the randomized s gives standard instance-level ϵ-DP for the whole record (Section 4.2).

3 Related Work

Latent-space DP. LATENT (Arachchige et al., 2020), ARDEN (Wang et al., 2018), and DP-VAE (Weggenmann et al., 2022) perturb every latent coordinate regardless of its role. DP representation learning (Sander et al., 2024) and a representation hand-designed for a Gaussian mechanism (Reimherr et al., 2025) likewise do not explicitly designate a protected coordinate for the release mechanism. **Directional mechanisms.** Metric DP scales protection with distance (Chatzikokolakis et al., 2013); von Mises–Fisher and Purkayastha mechanisms (Weggenmann & Kerschbaum, 2021) and directional gradient noise (Faustini et al., 2022) protect the angular coordinate, the inverse of our scoped formulation. **Selective protection.** Pufferfish/Blowfish (Kifer & Machanavajjhala, 2014; He et al., 2014), one-sided DP (Kotsogiannis et al., 2020), label DP (Chaudhuri & Hsu, 2011; Ghazi et al., 2021), and sensitive privacy for anomaly identification (Asif et al., 2019) define what may remain public but take the data structure as given; ORBIT supplies the representation-side step. **Private release under imbalance.** DP-GAN (Xie et al., 2018), PATE-GAN (Jordon et al., 2019), DP-EM (Park et al., 2017), and P3GM (Takagi et al., 2021) protect generator training rather than

each released record. Synthetic data protects outliers least (Stadler et al., 2022), and DP training harms underrepresented groups (Bagdasaryan et al., 2019; Feldman, 2020; Rosenblatt et al., 2025). Our imbalance limit follows from the hypothesis-testing view of DP (Dwork & Roth, 2014); we use it to measure how far a release is from what any per-record mechanism can achieve. **Norm and direction.** Deep SVDD scores anomalies by radius (Ruff et al., 2018), angular margins concentrate class directions (Deng et al., 2019), and the feature norm carries quality and out-of-distribution signal (Meng et al., 2021; Park et al., 2023); none of these addresses private release.

4 THE ORBIT FRAMEWORK

ORBIT instantiates the protected-coordinate formulation of Section 2. Latent-space DP usually adapts the mechanism to a fixed representation. ORBIT instead shapes the representation *for* the mechanism, in three phases: dual-shell shaping, radial perturbation, and release, with a direction-only audit when the direction is published (Section 6.4). Calibrated noise enters *only* the radial degree of freedom, which contains the class margin and the within-class radial variation designated as privacy-sensitive. Two design quantities tie the two sides together: the inter-shell margin $M = R_2 - R_1$ controls the crossover probability P_{cross} and the effective margin γ_{eff}, and the release window $[R_1, R_{max}]$ with μ_{max} sets the sensitivity Δh; their ratio $\Delta h/M$ is the budget penalty relative to the two-point release of Proposition 2 (Section 5). The full procedure is formalized in Algorithm 1 in Appendix A.

4.1 DUAL-SHELL REPRESENTATION SHAPING

The first phase shapes the latent space for two purposes. It places the class margin on the radial coordinate, whose within-class variation is designated as privacy-sensitive, so that the protected quantity is one scalar with class-separating structure. The projection of Section 4.2 bounds its range, while shaping makes this fixed range usable: on the evaluated representations, the projection changes the protected quantity for almost no record (Appendix G). Empirically, the directions within each class also become concentrated (Appendix H), which provides the premise of Assumption 1 for bounding directional leakage; whether that leakage is small in practice is audited separately (Section 6.4). Let $E_\theta : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^d$ be the encoder, anchored at the origin, and $\mathbb{I}_{out}(x) \in \{0, 1\}$ the anomaly indicator. Drawing inspiration from hyperspherical anomaly detection (Ruff et al., 2018) and large-margin learning (Deng et al., 2019), we define the *dual-shell margin loss* as a squared shell-targeting penalty,

$$\mathcal{L}_{margin} = (1 - \mathbb{I}_{out}(x)) \left(\|z\|_2 - R_1\right)^2 + \mathbb{I}_{out}(x) \left(\|z\|_2 - R_{out}\right)^2, \tag{2}$$

optimized jointly with a reconstruction loss through a curator-side decoder that is never released. Here, R_1 is the inner-shell radius and $[R_2, R_{max}]$ is the outer band, with $R_1 < R_2 < R_{max}$. The outlier target $R_{out} = (R_2 + R_{max})/2$ is the midpoint of this band; this ensures that anomalies settle strictly beyond the *vacuum margin* $[R_1, R_2]$ at convergence (the theoretical analysis requires only that outliers lie at or beyond R_2). We use a squared penalty (rather than a one-sided hinge): its unique per-sample optimum places each class *on* its target shell (the premise of the crossover analysis in Theorem 3), while the objective itself provides soft geometric guidance rather than a rigid projection (Appendix A).

4.2 UNIFIED DUAL-SHELL RADIAL PERTURBATION

The second phase imposes Definition 3 while leaving the direction untouched. Every latent vector first passes through a deterministic *feasible-domain projection* onto the radial band $[R_1, R_{max}]$,

$$\hat{z} = \Pi(z) = \mathrm{clip}\big(\|z\|_2, R_1, R_{max}\big)\, u(z), \qquad u(z) = z/\|z\|_2 \ (z \neq 0), \quad u(0) = u_0, \tag{3}$$

where $\mathrm{clip}(x, a, b) = \min(\max(x, a), b)$ and u_0 is a fixed public unit vector. This projection Π enforces $R_1 \leq \|\hat{z}\|_2 \leq R_{max}$ for every input, regardless of convergence, so the sensitivity bound requires no assumption about the encoder output. Specifically, Π preserves the direction of every $z \neq 0$ and maps $z = 0$ to u_0; its trigger rates are reported in Appendix G. The public direction is $\vec{u} = \hat{z}/\|\hat{z}\|_2$. The protected quantity of Section 2 is computed on the projected representation, $s = h(\hat{z}) = \|\hat{z}\|_2 + \tilde{\mu}(\hat{z})$. It is released through a one-dimensional Laplace mechanism and recombined with the public direction:

$$r_{priv} = h(\hat{z}) + \eta, \quad \eta \sim \mathrm{Laplace}(0, b), \qquad z_{priv} = \mathrm{clip}(r_{priv}, R_1, R_{max}) \cdot \vec{u}. \tag{4}$$

The location term $\tilde{\mu} \in [-\mu_{max}, \mu_{max}]$ is a clipped, label-free *gravity* term that pulls each record toward the nearer shell (Equation (6) in Appendix A); it is an implementation-level refinement that gives no measurable gain in our experiments, and $\mu_{max} = 0$ removes it. The projection onto $[R_1, R_{max}]$, together with the clipping of $\tilde{\mu}$, yields the finite sensitivity of Theorem 1. Π is a deterministic step inside the release mechanism, not a separate release, so its effect on s is included in this sensitivity, which Theorem 2 uses to calibrate the noise. The clip before recombination is post-processing and keeps the released direction equal to $\vec{u}$ (Appendix A). The *margin-integrity condition* $b \ll M/2$ preserves this shell topology in the presence of noise (Theorem 3).

Protected release and downstream use. *Radius-only release* publishes only r_{priv}. Because Theorem 1 bounds h over all feasible inputs, this is standard instance-level ϵ-DP for the whole record, with no public component. *Scoped per-record release* shares z_{priv} directly: when b is small relative to M, the margin absorbs radial noise and crossover is exponentially suppressed (Theorem 3), so a classifier trained on z_{priv} can use the protected radius as well as the public direction. *Fully synthetic release* fits class-conditional density estimators on the protected latents and samples balanced synthetic latents. A latent-space consumer uses them directly, so no decoder enters the privacy claim; an input-space release needs a public or DP-trained decoder G_ϕ for post-processing to apply. Like the scoped release, it is covered by Definition 3, not by whole-record DP.

5 Theoretical Analysis

We first state the limit that any per-record release faces under extreme imbalance and a release that attains it, against which Section 6 measures every mechanism; we then state the guarantees of ORBIT's radial release and, for the scoped release, the conditional bound on directional leakage (proofs in Appendix B).

The imbalance limit and its attainment. Let the evaluated records contain r normal records per anomaly. *Proposition 1 (Imbalance limit).* If each record is released by a mechanism that satisfies instance-level ϵ-DP with respect to the whole record, then for every detector applied to the released records and every decision threshold, the population rates satisfy $\mathrm{TPR} \leq e^\epsilon \mathrm{FPR}$. Hence the precision at every recall, and therefore the AUPRC, is at most $e^\epsilon/(e^\epsilon + r)$. In particular, precision above $1/2$ at any recall requires $\epsilon > \ln r$ (6.4 at 577:1), whatever the mechanism. The bound is a direct consequence of the definition (cf. the hypothesis-testing view of DP (Dwork & Roth, 2014)). It does not cover the scoped release, which declares the class public and publishes a direction that may encode it. *Proposition 2 (Two-point attainment).* If every normal record has protected value R_1, every anomaly has R_2, and the value is released with $\mathrm{Laplace}(0, M/\epsilon)$, i.e., with $\Delta h = M$, then $\mathrm{TPR} = e^\epsilon \mathrm{FPR}$ for every threshold $t \geq R_2$, so the precision equals the bound of Proposition 1 at every recall up to $1/2$. With $R_{max} = R_2$ and $\mu_{max} = 0$, Theorem 1 below gives $\Delta h = M$, and the dual-shell radius approximates this two-point release. A window of width $\Delta h > M$ gives $\mathrm{TPR}/\mathrm{FPR} = e^{\epsilon M/\Delta h}$, so it needs a privacy budget larger by the factor $\Delta h/M$.

Sensitivity and privacy. *Theorem 1 (Global sensitivity bound).* Under the projection of Equation (3) and the clipping $\tilde{\mu} \in [-\mu_{max}, \mu_{max}]$, the global sensitivity of $h(\hat{z}) = \|\hat{z}\|_2 + \tilde{\mu}(\hat{z})$ satisfies $\Delta h \leq (R_{max} - R_1) + 2\mu_{max}$: admissible norms lie in $[R_1, R_{max}]$ and the gravity term lies in a window of width $2\mu_{max}$, so Δh does not depend on d. *Theorem 2 (Radial ϵ-DP).* For any projected representation, the release $z_{priv} = \mathrm{clip}(r_{priv}, R_1, R_{max})\,\vec{u}$ with $r_{priv} = h(\hat{z}) + \mathrm{Laplace}(0, b)$ satisfies direction-public instance-level ϵ-DP (Definition 3) with respect to the protected quantity s whenever $b \geq \Delta h/\epsilon$: for a fixed direction $\vec{u}$, the released object is a scalar from $\mathrm{Laplace}(h(\hat{z}), b)$, whose density ratio between feasible inputs is at most $e^{\Delta h/b}$; the clip and the recombination with $\vec{u}$ follow by post-processing.

Crossover probability and its utility consequence. *Theorem 3 (Bounded crossover probability).* Let $M = R_2 - R_1$. For converged core representations at R_1 and converged anomalous representations at $R_{out} > R_2$, the probability that the noisy radius r_{priv} crosses to the opposite side of the vacuum margin satisfies

$$P_{cross} \leq \tfrac{1}{2} \exp(-M/b)\,, \tag{5}$$

independently of d: gravity places the expected radius of a converged record at or beyond the near edge of the margin ($h = R_1$ for core records, $h \geq R_2$ for anomalies), so crossing requires a Laplace deviation of at least M, whose tail mass is $\frac{1}{2}e^{-M/b}$ in either direction. For a non-converged record

whose gravity term is clipped, the expected radius may lie up to μ_{max} inside the margin, giving the weaker worst-case bound $\frac{1}{2}\exp(-(M-\mu_{max})/b)$. A union bound over the n released records turns this into an effective margin $\gamma_{eff} \geq M - b\ln\frac{n}{2\delta}$ with probability at least $1-\delta$, and hence into a standard margin-based generalization bound (Corollary 1, Appendix B); isotropic noise of scale b_{std} on all d coordinates instead erodes the margin at rate $\Theta(\sqrt{d}\, b_{std})$.

Directional concentration (scoped release). Let I be the identity of a record, Y its class (a deterministic function of I), and $U = \vec{u} = \hat{z}/\|\hat{z}\|_2 \in \mathbb{S}^{d-1}$ its released direction ($I \to X \to Z \to U$). *Assumption 1 (Approximate vMF convergence).* After Stage-1 shaping has converged, the direction of any identity i is well approximated by a shared class-conditional von Mises–Fisher law, $U \mid I{=}i \sim \mathrm{vMF}(\mu_{Y(i)}, \kappa_{Y(i)})$ with $\kappa_{Y(i)} \gg 1$ (cf. Deng et al., 2019). *Proposition 3 (Angular information compression).* Under the idealized limit of Assumption 1, $U \perp\!\!\!\perp I \mid Y$ and hence $I(U; I \mid Y) = 0$: if every identity in a class draws its direction from the same class law, the direction cannot single out the identity. *Lemma 1 (Bounded directional distinguishability).* If $I(U; I \mid Y) \leq \tau$, the average total-variation distinguishability between an individual-conditioned direction distribution and its class-conditional counterpart is at most $\sqrt{\tau/2}$ (by Pinsker's inequality; $\tau = 0$ in the idealized limit). Lemma 1 is scoped to identity and membership *conditional on the class*; whether real representations are in the low-τ regime is assessed empirically (Section 6.4, Appendix H).

6 Experimental Evaluation

We evaluate on three highly imbalanced benchmarks: Credit Card Fraud (284,807 records, 492 frauds, 577:1, d=32), a Covertype split with the rarest class as anomaly (500:1, 54 features), and imbalanced CIFAR-10 (airplane as anomaly, 100:1, d=128). We ask how close a private per-record release comes to the imbalance limit (RQ1), whether the crossover bound holds on learned representations (RQ2), how the margin affects utility at fixed sensitivity (RQ3), how utility changes with the imbalance ratio (RQ4), and what the public direction leaks (Section 6.4). **Setup and calibration.** On the tabular datasets, d=32, R_1=2, R_2=5 (M=3), R_{max}=8, μ_{max}=0.5, and R_{out}=6.5; Covertype reuses the Credit Card recipe unchanged, and the shells form as designed (normal $\|z\|_2 \approx 2.0$, anomaly $\approx$ 5.3). On CIFAR-10, a CNN encoder maps images to d=128 with R_1=8, R_2=14, R_{max}=20, and μ_{max}=2 ($\Delta h = 16$). Encoders are selected on a validation split carved from the training data, never on the test split. Tabular results and CIFAR-10 synthetic-release results use three encoder training seeds, CIFAR-10 per-record results use the deployed encoder, and per-record results average five noise repetitions per seed; data, architectures, and training details are in Appendix C. All radial mechanisms are calibrated from the design bound R_{max}, never from empirical maxima. Baselines (Appendix C). Full-dimensional **Standard Isotropic** latent DP works on the ball of radius R_{max} and is calibrated to the ℓ_1 sensitivity of the vector ($b = 2R_{max}\sqrt{d}/\epsilon$) for the common guarantee; the per-coordinate calibration $b_{std} = 2R_{max}/\epsilon$ of prior work serves as a favorable reference. We also run **LATENT** (Arachchige et al., 2020) (ϵ/d per coordinate), directional **VMF** DP (Weggenmann & Kerschbaum, 2021; Faustini et al., 2022), which leaves the norm public, **Radial** (radial Laplace noise on the same shaped representation without gravity, $b = R_{max}/\epsilon$), an **unshaped** control (the recipe without $\mathcal{L}_{margin}$), and the DP generators **DP-MoG** (Park et al., 2017) and **PATE-GAN** (Jordon et al., 2019), whose ϵ covers generator training. One-dimensional **supervised scores** (logistic regression, GBDT, and an MLP with the encoder's architecture) are clipped to a window tuned per ϵ on validation data, a tuning that is not privacy-accounted and favors them. **ORBIT** (*Dual-Shell* in figures) uses $\Delta h = (R_{max} - R_1) + 2\mu_{max} = 7.0$; for the radius-only release we also report the fixed, data-independent window $[R_1, R_2]$ with μ_{max}=0 ($\Delta h{=}M$; Proposition 2).

6.1 Per-Record Release: Approaching the Imbalance Limit

We first compare releases that satisfy standard instance-level ϵ-DP for the whole record (RQ1): ORBIT's radius alone, one-dimensional supervised scores, and full-dimensional latent DP with ℓ_1 calibration. Training and test records are released alike, and the consumer ranks the test records by the released value (Figure 1; full table in Appendix D). The mean AUPRC of every release lies below the bound of Proposition 1. Full-dimensional latent DP stays near the base rate up to ϵ=20 on all three datasets (Credit Card 0.004, Covertype 0.005, CIFAR-10 0.013). ORBIT's radius over the window $[R_1, R_2]$ reaches 0.12, 0.74, and 0.80 at $\epsilon = 5, 10, 20$ on Credit Card (bound at ϵ=5:

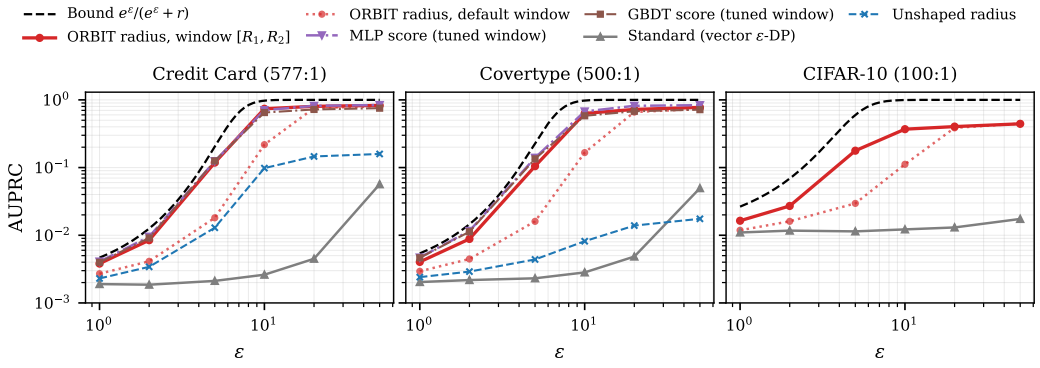


Figure 1: Per-record release under standard instance-level ϵ-DP for the whole record (no public component): AUPRC of ranking the released test records, against the bound of Proposition 1 (dashed). Tabular: mean over three seeds; CIFAR-10: deployed encoder. The supervised scores (MLP, GBDT) use a clip window tuned per ϵ; ORBIT's window $[R_1, R_2]$ is fixed. Full-dimensional latent DP stays near the base rate. The CIFAR-10 panel shows ORBIT and Standard only.

0.20), 0.11, 0.63, and 0.72 on Covertype, and 0.18, 0.37, and 0.40 on CIFAR-10. Above $\epsilon \approx 10$ it is limited by the radius's own non-private AUPRC (0.86, 0.83, and 0.56). With the default window (Δh=7), the same radius needs about twice the budget (0.22 at ϵ=10 and 0.76 at ϵ=20 on Credit Card), in line with the factor $\Delta h/M = 7/3$ of Proposition 2, because it releases many more normal records above R_2 (5.9% vs. 0.35% at ϵ=5; Appendix I).

We next test whether dimensional compression alone explains the gain. Released in the same way, the unshaped radius and unsupervised one-dimensional summaries (PCA, isolation forest, reconstruction error, a 1-D autoencoder) reach at most 0.43 on Credit Card and 0.03 on Covertype at ϵ=10 (Appendix D), whereas the coordinate shaped for the protected task signal reaches 0.74 and 0.63. The shaped radius is not a better scalar than a well-trained supervised score, however. With a window tuned per ϵ on validation, the MLP score is on par on Credit Card (0.71 at ϵ=10, 0.82 at ϵ=20) and higher on Covertype at $\epsilon \geq 10$ (0.67, 0.82); with the data-independent window $[0, 1]$ of its predicted probability it reaches 0.68 and 0.63 at ϵ=10. The gain over latent-space DP therefore comes from budgeting privacy on one bounded coordinate with class-separating structure, and the dual-shell radius achieves comparable utility while also yielding a representation (Section 6.2).

Proposition 1 also predicts that the attainable AUPRC falls as the imbalance ratio grows (RQ4). On pools drawn from the test split, which the encoder never saw, ORBIT's radius follows this on Credit Card: at ϵ=5 it reaches 0.76, 0.52, 0.38, 0.13, and 0.12 at 10:1, 50:1, 100:1, 500:1, and 577:1, against bounds of 0.94, 0.75, 0.60, 0.23, and 0.20, i.e., 58–81% of the bound (80–83% at ϵ=1; Appendix D). On CIFAR-10 the fraction is lower (e.g., 0.17 vs. 0.60 at 100:1), consistent with the radius's lower non-private AUPRC.

6.2 Representation-Level Release: Scoped and Synthetic

Publishing the direction as well (Section 2) places the release outside Proposition 1, because the class is declared public and the released direction may encode it. A classifier on z_{priv} then reaches 0.81 at ϵ=5 on Credit Card and 0.72 on Covertype, where Standard with the favorable per-coordinate calibration reaches 0.008 on both (Appendix D). This utility comes mainly from the public direction: releasing the direction alone gives 0.79 and 0.69, and the privatized radius adds to it only from $\epsilon \approx 5$ on (0.85 and 0.76 at ϵ=10; 0.86 and 0.75 at ϵ=20). The utility of the scoped release at small ϵ therefore reflects its scope rather than a better mechanism, and the public direction is audited for this reason (Section 6.4). The same holds across imbalance ratios (Figure 4).

For synthetic release on Credit Card and Covertype, every training record is released through the mechanism, class-conditional Gaussian mixtures are fitted on the released latents, and 2,000 synthetic latents per class are sampled, so no raw record enters downstream training. A logistic-regression classifier trained on the synthetic corpus is tested on the real test split, either in latent

space (Algorithm 1, Branch A; direction-public guarantee of Theorem 2) or after decoding through the curator's Stage-1 decoder. That decoder is not private, so decoded results measure decodability only (Section 4.2). Table 4 in Appendix D lists all values.

In latent space, ORBIT's release retains the non-private AUPRC at ϵ=1 on all three datasets (Credit Card 0.86 vs. 0.86, Covertype 0.77 vs. 0.76, CIFAR-10 0.46 vs. 0.45), so the learned representation remains useful when released as a latent representation, not only through its radial score. Standard Isotropic is comparable (0.84, 0.78, and 0.48), and all mechanisms are close to the non-private value at ϵ=10. Training directly on the released latents gives similar values (Appendix D), so density estimation does not explain them. The difference from per-record release is that the classifier is applied to clean test latents, so Proposition 1 does not constrain this protocol; Standard's favorable calibration also makes its values an optimistic reference. The mechanisms differ in decodability. Through the fixed decoder at ϵ=1 on Credit Card, Standard reaches 0.18, against 0.57 for ORBIT and 0.63 without privacy. Rescaling Standard's released latents into $[R_1, R_{max}]$ does not restore it (0.15), whereas VMF, which retains more of the direction, decodes at 0.53, consistent with directional information being important for decodability in this setting. On CIFAR-10, pixel-space TSTR collapses to the base rate for every method including Non-DP, so we make no pixel-level claim.

In this protocol, the representation changes utility more than the mechanism does. Without privacy, the unshaped encoder reaches only 0.65 on Credit Card, 0.17 on Covertype, and 0.06 on CIFAR-10, below every mechanism on the shaped encoder at ϵ=1. The unshaped encoder omits $\mathcal{L}_{margin}$, which uses the anomaly labels, so the gain reflects label-supervised shaping, not geometry alone. The DP generators, whose ϵ covers their training, do not improve on this: DP-MoG is below ORBIT on Credit Card (0.72 vs. 0.86) and far below it on Covertype (0.22 vs. 0.77), and PATE-GAN trained on the full imbalanced table reaches only 0.11 and 0.13 (realized $\epsilon \leq 1$).

6.3 Theorem Validation, Margin, and Dimension

The crossover bound holds on the learned representations (RQ2; Figure 5 in Appendix D). On Credit Card ($b \in [0.1, 5]$, $M \in \{1, 2, 3, 4\}$), the empirical crossover probability of the anomalies that satisfy the premise of Theorem 3 ($\|z\|_2 > R_2$) lies below $\frac{1}{2}e^{-M/b}$ in all 36 configurations. The bound is loose when the anomaly shell lies well beyond R_2 and approaches it as the shell nears R_2: the largest empirical-to-bound ratio is 0.59, 0.70, 0.84, and 0.98 for $M = 1, \ldots, 4$. On CIFAR-10 ($M \in \{2, 4, 6, 8\}$) the same holds for $M \leq 6$, and for M=8 the empirical curve coincides with the bound (deviations $\leq 6\%$), the case in which Theorem 3 holds with equality.

Theorem 3 and Corollary 1 also predict that, at a fixed noise scale, a wider margin lowers crossover and preserves the effective margin (RQ3). We retrain the Credit Card encoder for $M \in \{0, 0.5, 1, 2, 3, 4\}$ with R_{max}=8 fixed, so Δh=7 and $b = 7/\epsilon$ are the same for every M (three training seeds; Appendix D). Across the evaluated margins, the AUPRC of the released radius increases with M at every tested budget (Spearman $\rho = 1.0$ at each; e.g., $0.007 \to 0.025$ at ϵ=5 and $0.03 \to 0.31$ at ϵ=10) and the measured crossover decreases. Because Δh does not grow with M, we observe no interior optimum. Over the 30 (M, ϵ) configurations, the radius AUPRC tracks the measured crossover probability (Pearson $r = -0.94$ between their logarithms, partial $r = -0.71$ controlling for $\log \epsilon$; Figure 7a). Both come from the same released records, so this is a consistency check of the chain *margin or budget* $\to P_{cross} \to$ *utility*, not an independent test.

Radial perturbation also keeps the noise independent of d. On synthetic shell vectors at ϵ=10 (Figure 7b), ORBIT and Radial keep $\|z_{priv}\|_2/\|z\|_2 \approx 1$ and the direction unchanged at every $d \in \{8, \ldots, 512\}$, whereas Standard inflates the norm as $\Theta(\sqrt{d})$ (11.7× at d=512) and LATENT and VMF also degrade. Retraining the Credit Card encoder at $d \in \{8, \ldots, 128\}$ (Appendix D) shows that Standard, even with the favorable calibration, is near-random at $\epsilon \leq 5$ for every d, so its failure there does not require a high dimension. The dimension amplifies it at relaxed budgets: ranked by the released norm, its AUPRC at ϵ=20 falls from 0.07 at d=8 to 0.006 at d=128, whereas that of the radial release does not change with d.

6.4 Direction-Leakage Audit: the Operating Regime of the Public Coordinate

Publishing the direction is a policy choice; Lemma 1 bounds its conditional identity distinguishability only under a concentration condition, so we attack the released directions $\vec{u}$ on all datasets

Table 1: Direction-only membership attack (A1a), within-minority AUC (mean over 5 seeds; the strictest scope). "Chance" marks one-sided permutation $p > 0.05$ on all seeds. Identity linkage (A1b) top-1 accuracy stays close to the $1/K$ baseline on every dataset and scope (at most 0.135 at $K{=}10$ and 0.02 at $K{=}100$).

Representation	Credit Card	Covertype	CIFAR-10
ORBIT release $\vec{u}$	0.520 (chance)	0.508 (chance)	**0.713 (violated)**
Unshaped encoder (control)	0.474 (chance)	0.506 (chance)	0.488 (chance)
Isotropic-DP z (control)	0.445 (chance)	0.513 (chance)	0.458 (chance)

with adversaries observing only $\vec{u}$. The attacks score membership by maximum cosine similarity with disjoint reference and evaluation halves (A1a), rank $K \in \{10, 100, 1000\}$ linkage candidates by cosine similarity (A1b), and probe $\vec{u}$ for the class and, on Credit Card, the within-class amount decile (A3). As controls, we run the same attacks against an unshaped (reconstruction-only) encoder and against isotropic-DP representations at $\epsilon{=}1$. An attack is *consistent with chance* if its AUC is not above the permutation null on any of five attack seeds (one-sided, 1,000 resamples, $p > 0.05$).

On both tabular datasets, the released direction is consistent with chance on every membership scope (Table 1), and on Covertype this holds for all three encoder seeds (within-minority AUC 0.508, 0.530, 0.515). CIFAR-10 marks the failure regime. Its training minority holds only 432 images, and the within-minority membership attack reaches AUC 0.713, above chance on every seed, which suggests partial memorization of this small minority. In a checkpoint sweep, the leak is 0.49–0.62 before the anomaly shell forms and 0.68 at convergence, so in this sweep stopping early lowers the leak only by giving up the shell structure. The unshaped control does not leak (0.488). Identity linkage stays close to the $1/K$ baseline even here, although membership is detectable in aggregate.

Attribute leakage lies outside the scope of Lemma 1. On Credit Card, the released direction predicts the within-class amount decile above chance (probe accuracy 0.42 minority, 0.70 majority, vs. 0.10 chance). This is below the unshaped control ($0.48/0.82$) and falls to near chance under isotropic DP, but it remains measurable. As drawn in Section 2, attribute inference is outside the formal guarantee, and deployments in which coarse attributes are sensitive must handle it separately.

7 Discussion and Conclusion

We instantiate the framework on anomaly detection because rare-class geometry is especially vulnerable to privacy-induced distortion, and because within-class atypicality is a natural graded, individually sensitive quantity. Representations in which such a quantity can be confined to one coordinate may admit analogous constructions; our claims are limited to the evaluated settings.

Limitations. (i) Shaping requires anomaly labels or trusted indicators. (ii) With 432 training anomalies (CIFAR-10), the direction leaks minority membership (AUC 0.71), and in our checkpoint sweep no converged checkpoint passes the audit. (iii) Attribute leakage falls outside the formal scope of Lemma 1, and gravity gives no measurable gain. (iv) ORBIT's radius is on par with, not better than, a supervised score released in one dimension, and no class-protecting per-record release reaches precision $1/2$ unless $\epsilon > \ln r$. (v) In latent-space synthetic release, ORBIT's advantage over full-dimensional DP is limited to decodability at $\epsilon{=}1$ on Credit Card. (vi) Decoded results use the curator's non-private decoder; an end-to-end private input-space release requires a public or DP-trained decoder.

Conclusion. Under extreme imbalance, per-record private release of a representation faces a limit that no mechanism escapes: whenever the class is protected, precision is at most $e^{\epsilon}/(e^{\epsilon}{+}r)$. ORBIT shapes the representation so that the radial coordinate carries class-separating structure, and it releases only that coordinate, clipped to a fixed range. The released radius reaches 58–76% of the limit at $\epsilon = 5$–10 on Credit Card, while full-dimensional latent DP and unsupervised one-dimensional summaries stay far below it. Publishing the direction as well lies outside the limit, because it declares the class public, and we audit what the direction leaks. These results support allocating privacy according to the functional roles of the representation: protect the coordinate that carries the designated privacy-sensitive quantity, and release the public coordinate only after a leakage audit.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) for the following tasks with required disclosure: helping develop the conceptual framework and refine hypotheses (in particular the framing of the direction-public privacy model and its scope); designing and providing feedback on the research methodology and experiments (e.g., identifying mismatches between the stated and the executed evaluation protocols); implementing methods (experiment pipelines and the DP-MoG baseline, under author-specified protocols); cleaning and reformatting datasets (the data preparation scripts); interpreting results (drafting interpretations of experimental outcomes, which the authors checked against the data); and formulating mathematical claims, providing ingredients of their proofs, and assisting in the writing of proofs (the proofs in Appendix B, the worst-case remark for non-converged representations under Theorem 3, and the derivation of Corollary 1), all of which the authors verified by hand. We did not use generative AI tools to translate the manuscript, which was written in English; we used them only to translate drafts and related papers into our native language for reading. Generating synthetic datasets with AI tools and qualitative or thematic data analysis are not applicable to this work; the synthetic data in this paper are produced by the proposed method itself. Additionally, we used generative AI tools to edit the paper for readability, draft parts of the paper, suggest its structure under the page limit, brainstorm, modify figures, edit code, and search for related literature (candidates were then read, selected, and checked against their sources by the authors). Every number reported in the paper comes from an executed run with saved result files; no number is taken from model output without a corresponding run. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a privacy mechanism for sharing learned representations and uses only public benchmark datasets (Credit Card Fraud Detection, whose features apart from time and amount are anonymized principal components; UCI Covertype; CIFAR-10); it involves no human subjects and no new data collection. The main ethical risk lies in what ORBIT deliberately leaves public. Its formal guarantee covers only the protected radial quantity. Specifically, our audit shows that the public direction leaks membership when the minority class is very small (432 images on CIFAR-10) and carries residual within-class attribute signal (the transaction-amount decile on Credit Card). We report these failure regimes rather than hide them, scope every claim to the evaluated settings, and recommend that practitioners run direction-only audits on their own data before release and treat the direction as sensitive whenever coarse attributes carry sensitive information. The method is intended to reduce the disproportionate harm that privacy noise causes minority groups; it is not a substitute for protecting the encoder or the raw data, which must remain private.

REPRODUCIBILITY STATEMENT

Algorithm 1 (Appendix A) gives the complete procedure; all geometric parameters, noise calibrations, baseline settings, data construction, architectures, and training recipes are listed in Section 6 and Appendix C. Every assumption is stated explicitly, and complete proofs are in Appendix B. The evaluation protocols, including the decoder used in the decoded TSTR results, the semi-synthetic protocol, and the attack definitions and decision rule of the leakage audit, are described in Sections 6.2 and 6.4, as well as Appendix F. All datasets are public; the Covertype split and the CIFAR-10 imbalance are constructed as described in Section 6 and Appendix C. Runs use fixed random seeds (encoder seeds 42, 43, and 44; data split seed 42), and each trained encoder is shipped with its training configuration. The supplementary material contains anonymized code, the trained encoders, and all result files, together with a script that recomputes the reported numbers from these files; the code will be released publicly upon publication.

REFERENCES

Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7):5827–5842, 2020.

Hafiz Asif, Periklis A. Papakonstantinou, and Jaideep Vaidya. How to accurately and privately identify anomalies. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 719–736, 2019.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 15453–15462, 2019.

Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies Symposium (PETS)*, pp. 82–102, 2013.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pp. 155–186, 2011.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017.

Pedro Faustini, Natasha Fernandes, Shakila Tonni, Annabelle McIver, and Mark Dras. Directional privacy for deep learning. *arXiv preprint arXiv:2211.04686*, 2022.

Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 954–959, 2020.

Georgi Ganev, Meenatchi Sundaram Muthu Selva Annamalai, and Emiliano De Cristofaro. The elusive pursuit of reproducing PATE-GAN: Benchmarking, auditing, debugging. *Transactions on Machine Learning Research*, 2025.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 27131–27145, 2021.

Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: Tuning privacy-utility tradeoffs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 1447–1458, 2014.

Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies*, 2021(2):348–368, 2021.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations (ICLR)*, 2019.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning (ICML)*, volume 37 of *PMLR*, pp. 1376–1385, 2015.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1):1–36, 2014.

Ios Kotsogiannis, Stelios Doudalis, Sam Haney, Ashwin Machanavajjhala, and Sharad Mehrotra. One-sided differential privacy. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 493–504, 2020.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14225–14234, 2021.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh. Understanding the feature norm for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1557–1567, 2023.

Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. DP-EM: Differentially private expectation maximization. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 896–904, 2017.

Iker Perez, Jason Wong, Piotr Skalski, Stuart Burrell, Richard Mortier, Derek McAuley, and David Sutton. Locally differentially private embedding models in distributed fraud prevention systems. *arXiv preprint arXiv:2401.02450*, 2024.

Matthew Reimherr, Carlos Soto, Mark Shriver, and Aleksandra Slavković. Gaussian differentially private human faces under a face radial curve representation. In *International Conference on Learning Representations (ICLR)*, 2025.

Lucas Rosenblatt, Yuliia Lut, Ethan Turok, Marco Avella Medina, and Rachel Cummings. Differential privacy under class imbalance: Methods and empirical insights. In *International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 52065–52109, 2025.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Tom Sander, Yaodong Yu, Maziar Sanjabi, Alain Oliviero Durmus, Yi Ma, Kamalika Chaudhuri, and Chuan Guo. Differentially private representation learning via image captioning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 43255–43275, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium*, pp. 1451–1468, 2022.

Shun Takagi, Tsubasa Takahashi, Yang Cao, and Masatoshi Yoshikawa. P3GM: Private high-dimensional data release via privacy preserving phased generative model. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 169–180, 2021.

Ji Wang, Jianguo Zhang, Weidong Bao, Xiaomin Zhu, Bokai Cao, and Philip S Yu. Not just privacy: Improving performance of private deep learning in mobile cloud. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2407–2416, 2018.

Benjamin Weggenmann and Florian Kerschbaum. Differential privacy for directional data. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1205–1222, 2021.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. DP-VAE: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022 (WWW)*, pp. 721–731, 2022.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

A Algorithm and Implementation Details

Why a squared shell-targeting penalty rather than a one-sided hinge. A one-sided hinge such as $\max(0, \|z\|_2 - R_1)$ is minimized by *any* norm in $[0, R_1]$, leaving the equilibrium location of normal samples underdetermined: the entire interior of the ball is optimal. The projection would then map these records onto R_1, so the lower clip rather than the encoder would set the protected quantity of many records. The squared penalty of Equation (2) places its unique optimum at $\|z\|_2 = R_1$ for core samples and $\|z\|_2 = R_{out}$ for outliers, so that each class settles onto its target shell at convergence, while the objective itself provides soft geometric guidance (optimized jointly with the reconstruction loss) rather than a rigid projection: the network seeks a geometric equilibrium between representation fidelity and spatial isolation without shattering the data manifold. Shell convergence and the resulting angular concentration are verified in Appendix H.

Feasible-domain projection. Formally, the projection of Equation (3) is

$$\hat{z} = \begin{cases} R_1\, u_0, & z = 0, \\ R_1\, z/\|z\|_2, & 0 < \|z\|_2 < R_1, \\ z, & R_1 \le \|z\|_2 \le R_{max}, \\ R_{max}\, z/\|z\|_2, & \|z\|_2 > R_{max}. \end{cases}$$

The two clips play different roles. Consider the counterfactual mechanism without the upper clip, in which $\hat{z} = z$ for all $\|z\|_2 \ge R_1$. For $\|z\|_2 > R_2 + \mu_{max}$, its protected quantity is $s = h(\hat{z}) = \|z\|_2 - \mu_{max}$, which grows without bound, so no finite Δh holds over $\mathbb{R}^d$. The upper clip is therefore required for any finite bound over all inputs. The lower clip is not required for finiteness. It tightens the bound obtained by the argument of Theorem 1 from $R_{max} + 2\mu_{max}$ to $(R_{max} - R_1) + 2\mu_{max}$. The direction $z/\|z\|_2$ is undefined at $z = 0$. Mapping this case to the fixed public direction u_0 makes Π a deterministic map defined on all of $\mathbb{R}^d$. This projection Π is a deterministic step inside the release mechanism, not a separate release. Its effect on the protected quantity s is included in the sensitivity of $h \circ \Pi$ that Theorem 2 uses. The guarantee is stated for fixed values of the mechanism parameters R_1, R_2, R_{max}, μ_{max}, and u_0. On the evaluated representations, the upper clip changes at most three records per Credit Card encoder and none on CIFAR-10, and $z = 0$ never occurs (Appendix G).

Gravity term. With $R_{mid} = (R_1 + R_2)/2$, the location term of Section 4.2 is

$$\mu(z) = \begin{cases} R_1 - \|z\|_2, & \|z\|_2 \le R_{mid} \\ R_2 - \|z\|_2, & \|z\|_2 > R_{mid}, \end{cases} \qquad \tilde{\mu}(z) = \max\big(-\mu_{max}, \min(\mu_{max}, \mu(z))\big). \quad (6)$$

It stabilizes radial positions around the target shells without accessing labels at release time; the selective-protection principle rests on the dual-shell geometry alone, and gravity gives no measurable gain over the gravity-free radial release in our experiments.

Direction preservation. The Laplace support is unbounded, so $r_{priv} < 0$ is possible when $\eta < -h(\hat{z})$; its probability is $\frac{1}{2}\exp(-h(\hat{z})/b) \le \frac{1}{2}\exp(-(R_1 - \mu_{max})/b)$, since $h(\hat{z}) \ge R_1 - \mu_{max}$ after projection and clipping. For a converged normal record ($h = R_1$) it is 0.38 at ϵ=1 and 0.03 at ϵ=10 on the tabular datasets ($b = 7/\epsilon$), and 0.30 and 0.003 on CIFAR-10 ($b = 16/\epsilon$). Recombining such a value as $r_{priv}\vec{u}$ would invert the direction. The vector release therefore clips r_{priv} to $[R_1, R_{max}]$ before recombination (Equation (4)), so the released direction always equals $\vec{u}$. The clip is a deterministic function of the released scalar, so the guarantee is unaffected. Without it, the

inverted directions remove most of the utility of the vector release at small ϵ: on Credit Card at ϵ=1, the scoped per-record LR-AUPRC falls from 0.65 to 0.04 and the latent-space TSTR AUPRC from 0.86 to 0.76. The radius-only release publishes r_{priv} itself.

Margin-integrity condition. The clipping $\tilde{\mu} \in [-\mu_{max}, \mu_{max}]$ (Equation 6) together with the projection onto $[R_1, R_{max}]$ (Equation 3) yields the finite sensitivity of Theorem 1. Separately, to preserve the dual-shell topology under perturbation, the noise scale is constrained by the margin width,

$$b \ll \frac{R_2 - R_1}{2}, \tag{7}$$

under which the probability of topological crossover, that is, an inner-shell record leaking into the outer band or vice versa, decays exponentially (Theorem 3).

Algorithm. Algorithm 1 gives the end-to-end procedure. Stage 1 minimizes the reconstruction loss jointly with $\mathcal{L}_{margin}$, guiding normal instances onto the R_1 shell and anomalies onto R_{out}, carving out the vacuum margin $[R_1, R_2]$. Stage 2 applies the unified radial perturbation to every latent vector, protecting the radial quantity while keeping the direction fixed. Stage 3 fits class-conditional density estimators on the protected latents and supports latent-space release or input-space synthesis with a privacy-compliant decoder.

B PROOFS

Proof of Proposition 1. Fix the detector, including any model trained on other released records, and let S be its positive region at a given threshold. For any anomaly x and any normal record x', instance-level ϵ-DP gives $\Pr[\mathcal{M}(x) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(x') \in S]$. Averaging over the anomalies on the left and over the normal records on the right gives $\mathrm{TPR} \leq e^{\epsilon}\,\mathrm{FPR}$ for the population rates (probabilities under the mechanism's randomness, averaged over the records). With r normal records per anomaly, the precision is $\mathrm{TPR}/(\mathrm{TPR} + r\,\mathrm{FPR}) \leq \mathrm{TPR}/(\mathrm{TPR} + re^{-\epsilon}\mathrm{TPR}) = e^{\epsilon}/(e^{\epsilon} + r)$. The AUPRC (average precision) is a weighted average of precisions at the attained recalls, so it obeys the same bound. An AUPRC measured on one test split and one noise draw estimates this population quantity and can exceed the bound by sampling noise (Table 3). The constraint $\mathrm{TPR} \leq e^{\epsilon}\,\mathrm{FPR}$ follows from the hypothesis-testing characterization of ϵ-DP (Kairouz et al., 2015), and an analogous argument bounds the positive predictive value of membership inference under a skewed prior (Jayaraman et al., 2021). ■

Proof of Proposition 2. With $b = M/\epsilon$ and $\eta \sim \mathrm{Laplace}(0, b)$, for $t \geq R_2$ we have $\mathrm{TPR} = \Pr[R_2 + \eta > t] = \frac{1}{2}e^{-(t-R_2)/b}$ and $\mathrm{FPR} = \Pr[R_1 + \eta > t] = \frac{1}{2}e^{-(t-R_1)/b}$, so $\mathrm{TPR}/\mathrm{FPR} = e^{M/b} = e^{\epsilon}$ and the precision equals $e^{\epsilon}/(e^{\epsilon} + r)$. As t decreases from ∞ to R_2, the recall covers $(0, 1/2]$. With a window of width $\Delta h > M$, $b = \Delta h/\epsilon$ and the same argument gives $\mathrm{TPR}/\mathrm{FPR} = e^{\epsilon M/\Delta h}$. ■

Proof of Theorem 1. The feasible-domain projection of Equation (3) is a deterministic step inside the release mechanism, applied identically to every record before noise is added. It is not released separately; its effect on the protected quantity s is accounted for by bounding the sensitivity of $h \circ \Pi$ below. After the projection, every admissible norm satisfies

$$R_1 \leq \|\hat{z}\|_2 \leq R_{max}, \tag{8}$$

where both bounds are enforced by the projection for every input. Then

$$\begin{aligned} \Delta h &= \max_{\hat{z}, \hat{z}'} \left| (\|\hat{z}\|_2 + \tilde{\mu}(\hat{z})) - (\|\hat{z}'\|_2 + \tilde{\mu}(\hat{z}')) \right| \\ &\leq \max_{\hat{z}, \hat{z}'} \left| \|\hat{z}\|_2 - \|\hat{z}'\|_2 \right| + \max_{\hat{z}, \hat{z}'} \left| \tilde{\mu}(\hat{z}) - \tilde{\mu}(\hat{z}') \right|. \end{aligned} \tag{9}$$

By Equation (8), the first term is at most $R_{max} - R_1$, and the clipping $\tilde{\mu} \in [-\mu_{max}, \mu_{max}]$ bounds the second by $2\mu_{max}$. ■

Proof of Theorem 2. For every fixed public direction $\vec{u}$, the mechanism releases a scalar $v \in \mathbb{R}$ drawn from $\mathrm{Laplace}(s, b)$ with $s = h(\hat{z})$ the protected radial quantity. Reconstructing $z_{priv} = \mathrm{clip}(v, R_1, R_{max})\,\vec{u}$ is deterministic given $\vec{u}$, so the privacy loss is determined solely by the density

Algorithm 1 ORBIT: Protected Representation Release and Optional Synthesis

Require: Dataset $\mathcal{D} = \mathcal{D}_{norm} \cup \mathcal{D}_{anom}$, Boundary radii $R_1 < R_2 < R_{max}$, Gravity bound μ_{max}, Privacy budget ϵ.

Ensure: Protected representations $\mathcal{Z}_{priv}$, with optional latent-space or input-space synthetic release.

// Stage 1: Dual-Shell Margin Shaping
Initialize latent Encoder E_θ and a curator-side Decoder G_ϕ^{shape} (a training-time auxiliary decoder that is never released).
repeat
  Sample a mini-batch $\mathcal{B}$ from raw dataset $\mathcal{D}$.
  Map instances to the latent space: $Z = E_\theta(\mathcal{B})$.
  Compute $\mathcal{L}_{margin}(Z; R_1, R_{out})$ to guide normal instances onto the R_1 shell and anomalies onto the outer-band target R_{out}.
  Compute reconstruction loss $\mathcal{L}_{rec}(\mathcal{B}, G_\phi^{shape}(Z))$.
  Update parameters θ, ϕ via gradient descent to minimize $\lambda \cdot \mathcal{L}_{margin} + \mathcal{L}_{rec}$.
until convergence criteria met
// Stage 2: Unified Dual-Shell Radial Perturbation
Extract all latent representations: $\mathcal{Z}_{all} = \{E_\theta(x) \mid x \in \mathcal{D}\}$.
Calibrate Laplace noise scale $b = \Delta h / \epsilon$.
Initialize the protected feature set $\mathcal{Z}_{priv} \leftarrow \emptyset$.
for each latent vector $z \in \mathcal{Z}_{all}$ **do**
  // Deterministic feasible-domain projection (Eq. 3; enforces $R_1 \le \|z\|_2 \le R_{max}$ before noise injection)
  if $z = 0$ **then**
    $z \leftarrow R_1 u_0$ {fixed public direction}
  else if $\|z\|_2 < R_1$ **then**
    $z \leftarrow R_1 \dfrac{z}{\|z\|_2}$
  else if $\|z\|_2 > R_{max}$ **then**
    $z \leftarrow R_{max} \dfrac{z}{\|z\|_2}$
  end if
  Record public direction $\vec{u} \leftarrow z / \|z\|_2$.
  Compute gravity $\tilde{\mu}(z)$ and the expected radius $h(z) = \|z\|_2 + \tilde{\mu}(z)$.
  Sample standard zero-mean Laplace noise: $\eta \sim Laplace(0, b)$.
  Compute the pre-clipping noisy radius $r_{priv} = h(z) + \eta$ and reconstruct $z_{priv} \leftarrow \text{clip}(r_{priv}, R_1, R_{max}) \cdot \vec{u}$ {post-processing}
  $\mathcal{Z}_{priv} \leftarrow \mathcal{Z}_{priv} \cup \{z_{priv}\}$.
end for
// Stage 3: Protected Density Modeling and Optional Synthetic Release
Partition the protected set by class: $\mathcal{Z}_{priv} = \mathcal{Z}_{priv}^{norm} \cup \mathcal{Z}_{priv}^{anom}$.
Fit class-conditional density estimators: P_{norm} over $\mathcal{Z}_{priv}^{norm}$ and P_{anom} over $\mathcal{Z}_{priv}^{anom}$.
Sample balanced synthetic latent vectors: $\mathcal{D}_{syn}^{latent} = Z_{syn}^{norm} \cup Z_{syn}^{anom}$.
// Branch A: Latent-Space Release (covered by Theorem 2 via post-processing)
Output $\mathcal{D}_{syn}^{latent}$ directly for downstream latent-space classification.
// Branch B: Input-Space Synthesis (Requires compliant decoder)
Decode latents using a public or DP-trained decoder G_ϕ^{DP}: $\mathcal{D}_{syn}^{input} = \{G_\phi^{DP}(z) \mid z \in \mathcal{D}_{syn}^{latent}\}$.

Return Protected latents $\mathcal{Z}_{priv}$ (and optionally $\mathcal{D}_{syn}^{latent}$ or $\mathcal{D}_{syn}^{input}$)

of v. For any two feasible protected values corresponding to the same $\vec{u}$,

$$\frac{p(v \mid \hat{z})}{p(v \mid \hat{z}')} = \frac{\exp(-|v - h(\hat{z})|/b)}{\exp(-|v - h(\hat{z}')|/b)} \leq \exp\left(\frac{|h(\hat{z}) - h(\hat{z}')|}{b}\right) \leq \exp\left(\frac{\Delta h}{b}\right). \tag{10}$$

Substituting $b = \Delta h/\epsilon$ yields e^{ϵ}. ■

Proof of Theorem 3. For a converged core representation, $\|\hat{z}\|_2 = R_1$ and $\mu(\hat{z}) = R_1 - \|\hat{z}\|_2 = 0$, so $h(\hat{z}) = R_1$. Crossing into the outer region requires $r_{priv} > R_2$, i.e. $\eta > M$, so $\Pr(r_{priv} > R_2) = \Pr(\eta > M) = \frac{1}{2}\exp(-M/b)$. For a converged anomalous representation, $\|\hat{z}\|_2 = R_{out} > R_2$ and $\mu(\hat{z}) = R_2 - R_{out}$, giving $h(\hat{z}) = R_2$ (the clipping window μ_{max} is chosen so that $R_{out} - R_2 \leq \mu_{max}$, or else $h(\hat{z}) > R_2$ and the bound is only more conservative). Crossing into the inner region requires $r_{priv} < R_1$, i.e. $\eta < -M$, so $\Pr(r_{priv} < R_1) = \frac{1}{2}\exp(-M/b)$. The same exponential bound holds in both directions. Crossover is defined on the pre-clipping noisy radius r_{priv}, which the radius-only release publishes directly. The vector release clips r_{priv} to $[R_1, R_{max}]$ only to preserve the direction. This post-processing does not alter the crossover event defined above, so the same bound applies to the vector release. The empirical crossover of Section 6.3 is measured on r_{priv}. ■

Remark (non-converged representations). For a representation whose gravity term is clipped at $\pm\mu_{max}$, the expected radius may lie up to μ_{max} inside the margin, and the same tail argument yields the worst-case bound $\frac{1}{2}\exp(-(M - \mu_{max})/b)$. Section 6.3 validates the converged bound on learned representations.

Corollary 1 (Margin-based utility consequence). For n independently released representations, a union bound gives $\Pr(\text{any crossover}) \leq \frac{n}{2}e^{-M/b}$. Hence, with probability at least $1 - \delta$, no representation crosses the margin whenever $M \geq b \ln \frac{n}{2\delta}$. Consequently, the effective margin seen by a downstream classifier satisfies $\gamma_{eff} \geq M - b \ln \frac{n}{2\delta}$, and a standard margin bound (Bartlett, 1998; Bartlett & Mendelson, 2002) gives

$$\text{err}_{gen} \leq \mathcal{O}\left(\frac{1}{\gamma_{eff}\sqrt{n}}\right) = \mathcal{O}\left(\frac{1}{\left(M - b \ln \frac{n}{2\delta}\right)\sqrt{n}}\right), \tag{11}$$

whereas isotropic noise of scale b_{std} on all d coordinates erodes the margin at rate $\Theta(\sqrt{d}\, b_{std})$. This is a consequence of Theorem 3 and standard large-margin bounds, not a new generalization result.

Derivation of Corollary 1. By the Laplace tail bound and a union bound over the n released representations, with probability at least $1 - \delta$ no pre-clipping noisy radius r_{priv} moves toward the opposite shell by more than $b \ln \frac{n}{2\delta}$, so the effective margin seen by the downstream classifier satisfies $\gamma_{eff} \geq M - b \ln \frac{n}{2\delta}$. Substituting γ_{eff} into a standard margin-based generalization bound (Bartlett, 1998; Bartlett & Mendelson, 2002) gives Equation (11). Under isotropic noise of scale b_{std} on all d coordinates, the L_2 displacement grows as $\Theta(\sqrt{d}\, b_{std})$, so the corresponding effective margin $M - \Theta(\sqrt{d}\, b_{std})$ shrinks with the dimension. ■

Proof of Proposition 3. Under the idealized shared-vMF assumption, the individual conditional density equals the shared class density, $p(u \mid I{=}i) = p(u \mid Y{=}Y(i))$, eliminating any dependence on the specific identity. This establishes $U \perp\!\!\!\perp I \mid Y$ and hence $I(U; I \mid Y) = 0$. In practice, the shared-vMF assumption is only approximate; the residual within-class identity dependence is assessed empirically (Section 6.4). ■

Proof of Lemma 1. By Pinsker's inequality,

$$D_{\text{TV}}(p(U \mid I, Y) \,\|\, p(U \mid Y)) \leq \sqrt{\tfrac{1}{2} D_{\text{KL}}(p(U \mid I, Y) \,\|\, p(U \mid Y))}.$$

Averaging over (I, Y) and applying Jensen's inequality,

$$\mathbb{E}_{I,Y}[D_{\text{TV}}(p(U \mid I, Y) \,\|\, p(U \mid Y))] \leq \sqrt{\tfrac{1}{2} I(U; I \mid Y)} \leq \sqrt{\tau/2}.$$

Under the idealized shared-vMF assumption, $\tau = 0$. For real representations, the empirical leakage audit evaluates whether direction-only attacks remain within this low-distinguishability regime. ■

Remark (scope). Lemma 1 is scoped to identity and membership information conditional on the class label; it does not assert that the direction is uninformative. The class is deliberately public,

and residual within-class attribute signal (e.g., coarse transaction-amount level) is outside the guarantee; Section 6.4 measures both separately. The formal zero-leakage result relies on the idealized shared-vMF assumption; residual dependence is not eliminated by the assumption itself and must be assessed empirically (Appendix H, Section 6.4).

C Baselines and Calibration Details

To ensure a fair comparison, the global sensitivity of all radial DP mechanisms is calibrated from the design bound R_{max} rather than empirical maxima. Standard Isotropic and Radial are calibrated to the ball of radius R_{max} and project every representation onto it ($z \mapsto z \min(1, R_{max}/\|z\|_2)$) before adding noise, so their stated guarantees hold for every input. The projection changes at most four records per shaped tabular encoder (maximum norm 8.29 on Credit Card and 9.96 on Covertype) and none on CIFAR-10 (Appendix G).

- **Non-DP (upper bound):** the original latent representations without perturbation.
- **Standard Isotropic DP:** independent Laplace noise on every latent dimension, $b_{std} = 2R_{max}/\epsilon$, which also perturbs the direction. This scale bounds the sensitivity of each coordinate ($|z_i| \le R_{max}$). The ℓ_1 sensitivity of the whole vector over the ball of radius R_{max} is $2R_{max}\sqrt{d}$, so the release satisfies $\sqrt{d}\,\epsilon$-DP for the vector rather than ϵ-DP. Standard therefore receives less noise than an ϵ-DP release of the full vector would require, which favors this baseline. For the common-guarantee comparison (Section 6.1) we therefore also run Standard with $b = 2R_{max}\sqrt{d}/\epsilon$, which satisfies ϵ-DP for the vector.
- **LATENT** (Arachchige et al., 2020): a full-domain randomization layer originally proposed under local DP, which discretizes each latent dimension and applies randomized response; under sequential composition, the budget is split as ϵ/d, the canonical high-dimensional failure mode of per-dimension budget splitting.
- **VMF Directional DP** (Faustini et al., 2022; Weggenmann & Kerschbaum, 2021): the conceptual mirror image of ORBIT. It treats the radial magnitude as public and protects the direction via von Mises–Fisher sampling, $\vec{u}_{priv} \sim \mathrm{vMF}(\vec{u}, \kappa)$ with $\kappa = \epsilon R_{max}$ (tangent-space Gaussian approximation), leaving $\|z\|_2$ intact. This is a metric-DP calibration, weaker than ϵ-DP for the direction.
- **Radial (ablation):** adheres to the direction-public assumption and perturbs only the radial magnitude of the same shaped representation, without gravity, $r_{priv} = \min(\|z\|_2, R_{max}) + \mathrm{Laplace}(0, R_{max}/\epsilon)$, with the same clip of the released radius to $[R_1, R_{max}]$ before recombination as ORBIT.
- **Unshaped control:** the same encoder architecture and training recipe without $\mathcal{L}_{margin}$ (reconstruction loss only), used to separate the effect of shaping from that of the mechanism.
- **DP generative model** (TSTR): **DP-MoG**, a per-class diagonal Gaussian on the raw features privatized through clipped sufficient statistics, in the spirit of DP-EM (Park et al., 2017). Its ϵ covers the entire generator training, whereas ORBIT's covers the released radius.
- **PATE-GAN** (Jordon et al., 2019): synthcity implementation with default settings (teacher-vote noise $\lambda = 10^{-3}$, at most 200 iterations), trained on the full imbalanced table or on the minority partition (semi-synthetic protocol). Its ϵ is a stopping criterion of the moments accountant; on the full tables, training reaches the iteration cap before ϵ=1 is spent, so the ϵ=1 and ϵ=10 runs coincide. An audit of six PATE-GAN implementations, including synthcity's, found that none reproduces the originally reported utility (Ganev et al., 2025).
- **One-dimensional supervised scores** (common guarantee): the logit of a class-balanced logistic regression, of a histogram gradient-boosting classifier, and of an MLP with the encoder's architecture and a linear head (100 epochs; positive-class weight in $\{1, \sqrt{r}, r\}$ and epoch selected by validation AUPRC), all trained on the fit split. Each logit is clipped to a window $[a, b]$ chosen per ϵ among 16 quantile-based candidates by validation AUPRC and released as $\mathrm{clip}(s, a, b) + \mathrm{Laplace}(0, (b-a)/\epsilon)$. Table 2 also reports the MLP's predicted probability released with its data-independent window $[0, 1]$. The window selection is not privacy-accounted, which favors these baselines.

On the tabular datasets, ORBIT sets $\Delta h = (R_{max} - R_1) + 2\mu_{max} = 7.0$ ($b = 7/\epsilon$), injecting 12.5% less noise than Radial ($b = 8/\epsilon$) at the same budget. On CIFAR-10 (R_1=8, R_2=14, R_{max}=20, μ_{max}=2), ORBIT yields $\Delta h = 16$, 20% less noise than Radial ($b = 20/\epsilon$). The CIFAR-10 encoder is trained with random-crop and horizontal-flip augmentation and AdamW weight decay 5×10^{-4} (margin weight $\lambda = 50$, 80 epochs, last-epoch model). This recipe was selected on a validation split carved from the training data; the test split is not used for model selection. The tabular encoders (Credit Card and Covertype) are trained for 100 epochs with a cosine learning-rate schedule and margin weight $\lambda = 5$, and the model with the lowest validation loss is kept (validation: 20% of the training split; reconstruction loss only for the unshaped control). Tabular results use three encoder training seeds (42, 43, 44).

Data and implementation details. Credit Card uses V1–V28 and the transaction amount, standardized over the whole file without labels (the time field is dropped). Covertype uses its 54 features, standardized over the constructed dataset; the rarest cover type (class 4, 2,747 records) is the anomaly class, subsampled uniformly (seed 42) to 1,156 records, one per 500 of the 578,265 records of the other six types, for 579,421 records in total. The CIFAR-10 pool contains the 54,000 non-airplane images of the training and test sets and 540 randomly subsampled airplane images (100:1). Every dataset is split 80/20 into training and test sets (stratified, seed 42), and 20% of the training split is held out for model selection. On CIFAR-10 this split of the pooled images replaces the official train/test split; on Credit Card the test split has 580 normal records per anomaly. The tabular encoder is an MLP with hidden layers of 128 and 64 units (ReLU) and a mirrored decoder, trained with Adam (learning rate 10^{-3}, batch size 512) on the mean-squared reconstruction loss plus $\lambda\mathcal{L}_{margin}$. The CIFAR-10 encoder has four 3×3 convolutional blocks (32, 64, 128, and 256 channels, each with batch normalization, ReLU, and 2×2 max pooling) followed by fully connected layers of 256 and 128 units, with a transposed-convolution decoder (batch size 256, learning rate 10^{-3}). The downstream classifier is a class-balanced logistic regression on standardized inputs (scikit-learn). Per-record results use five noise repetitions per encoder seed (three for the margin and dimension sweeps). The fully synthetic pipeline fits full-covariance Gaussian mixtures per class, with the number of components chosen by BIC from $\{1, 2, 4, 8\}$, samples 2,000 latents per class (CIFAR-10: diagonal covariance, 432 latents per class), and uses ten repetitions per encoder seed. DP-MoG clips raw feature vectors to norm $C = 10$ and splits ϵ as $0.2/0.4/0.4$ over the noisy count, sum, and sum of squares. Tabular experiments run on a CPU; the CIFAR-10 encoders are trained on a single 12 GB GPU. Software: PyTorch 2.12, scikit-learn 1.8–1.9, NumPy 2.4, SciPy 1.17, and umap-learn 0.5; PATE-GAN uses synthcity 0.2.12 (with PyTorch 2.2).

D Additional Empirical Results

UMAP view of the scoped release. Figure 2 shows 1,200 random normal and all 98 anomalous test records of the deployed Credit Card encoder under the scoped per-record release at ϵ=5, with one UMAP fit per panel. Standard Isotropic removes the structure of the embedding, and the anomalies scatter across the normal mass. ORBIT's release keeps most of the non-private structure, because the direction is released unchanged, and most anomalies stay grouped. This is a view of the scoped release, whose utility comes mainly from the public direction (Section 6.1); the embedding is nonlinear and does not show how much a linear classifier recovers.

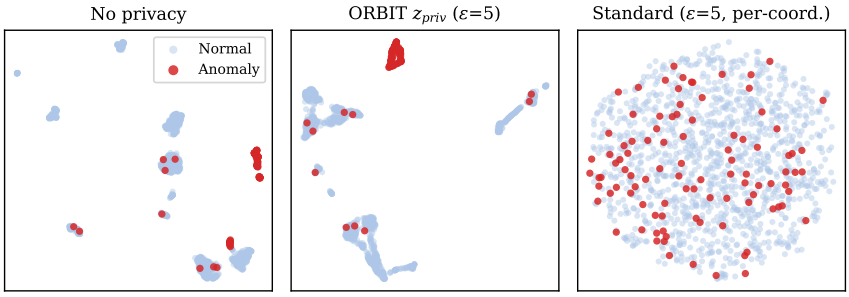


Figure 2: UMAP of Credit Card test latents under the scoped per-record release ($\epsilon = 5$, deployed encoder): no privacy (left), ORBIT z_{priv} with the default window (middle), and Standard with the per-coordinate calibration (right).

Common guarantee, full table. Table 2 lists the values behind Figure 1.

Table 2: Per-record release under standard instance-level ϵ-DP for the whole record: AUPRC of ranking the released test records (tabular: mean over three seeds; CIFAR-10: deployed encoder, five repetitions). r is the test-split imbalance. The last four rows of each tabular block are unsupervised one-dimensional summaries. †Window tuned per ϵ on validation (not privacy-accounted). *Per-coordinate calibration $b_{std} = 2R_{max}/\epsilon$, which satisfies only $\sqrt{d}\,\epsilon$-DP for the vector; best of Norm-AUPRC and LR-AUPRC.

Method $\epsilon =$	2	5	10	20	50
Credit Card ($r = 580$)					
Bound $e^{\epsilon}/(e^{\epsilon}+r)$	0.013	0.204	0.974	1.000	1.000
ORBIT radius, $[R_1, R_2]$ (Δh=M)	0.008	0.118	0.739	0.803	0.828
ORBIT radius, default (Δh=7)	0.004	0.018	0.218	0.756	0.806
MLP score†	0.010	0.124	0.705	0.817	0.839
MLP score, window $[0, 1]$	0.007	0.087	0.682	0.803	0.824
GBDT score†	0.009	0.125	0.650	0.722	0.754
LR score†	0.009	0.122	0.652	0.704	0.732
Standard (ℓ_1 calibration)	0.002	0.002	0.003	0.004	0.057
Standard (per-coordinate)*	0.003	0.008	0.083	0.590	0.818
Unshaped radius	0.003	0.013	0.098	0.146	0.159
PCA, first component†	0.003	0.006	0.009	0.010	0.009
Isolation forest†	0.007	0.038	0.104	0.133	0.149
Reconstruction error†	0.008	0.098	0.414	0.480	0.491
1-D autoencoder†	0.007	0.081	0.430	0.449	0.454
Covertype ($r = 501$)					
Bound $e^{\epsilon}/(e^{\epsilon}+r)$	0.015	0.229	0.978	1.000	1.000
ORBIT radius, $[R_1, R_2]$ (Δh=M)	0.009	0.105	0.631	0.723	0.775
ORBIT radius, default (Δh=7)	0.004	0.016	0.166	0.660	0.735
MLP score†	0.012	0.140	0.674	0.817	0.835
MLP score, window $[0, 1]$	0.008	0.096	0.626	0.735	0.782
GBDT score†	0.011	0.134	0.582	0.679	0.717
LR score†	0.010	0.079	0.200	0.245	0.264
Standard (ℓ_1 calibration)	0.002	0.002	0.003	0.005	0.050
Standard (per-coordinate)*	0.003	0.008	0.078	0.469	0.708
Unshaped radius	0.003	0.004	0.008	0.014	0.017
PCA, first component†	0.007	0.015	0.017	0.018	0.018
Isolation forest†	0.007	0.015	0.018	0.019	0.020
Reconstruction error†	0.006	0.013	0.017	0.017	0.018
1-D autoencoder†	0.008	0.019	0.022	0.022	0.024
CIFAR-10 ($r = 100$)					
Bound $e^{\epsilon}/(e^{\epsilon}+r)$	0.069	0.597	0.995	1.000	1.000
ORBIT radius, $[R_1, R_2]$ (Δh=M)	0.027	0.178	0.369	0.403	0.443
ORBIT radius, default (Δh=16)	0.016	0.030	0.112	0.386	0.446
Standard (ℓ_1 calibration)	0.012	0.011	0.012	0.013	0.017

Generic one-dimensional summaries. Table 2 also releases unsupervised one-dimensional summaries with the protocol of the supervised scores (same splits, window grid, and per-ϵ window tuning on validation, which favors them): the first principal component of the standardized features, an isolation-forest score, the reconstruction error of the unshaped encoder, and a 1-D bottleneck autoencoder trained with the unshaped recipe (the sign of the principal component and of the autoencoder output is chosen on validation). On Covertype none exceeds 0.03 at any budget. On Credit Card, fraud is partly visible without labels: the reconstruction error and the 1-D autoencoder reach 0.41 and 0.43 at ϵ=10 (mean over three seeds), the latter with 0.64 on two seeds and 0.01 on the third. The shaped radius reaches 0.74 and 0.63 with a fixed window.

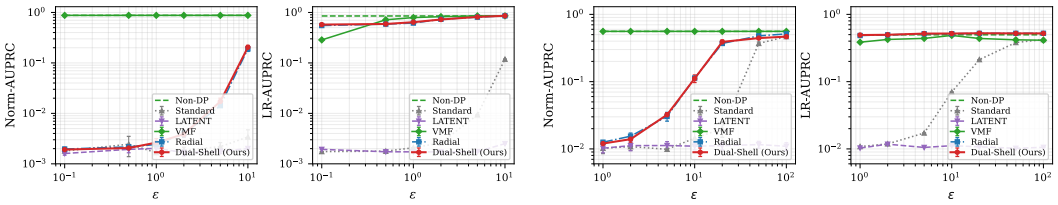
Imbalance under the common guarantee (RQ4). Table 3 reports the radius-only release (window $[R_1, R_2]$) across imbalance ratios. Each of 20 random draws forms a pool from the test split only, so no evaluated record was used to train or select the encoder: all test anomalies and r normal test records per anomaly (fewer anomalies when the test normals do not suffice, e.g., 21 and 10 on CIFAR-10 at 500:1 and 1000:1; "577:1" is the whole test split). The bound concerns population rates, so individual draws can exceed it by sampling noise when a release is close to it; the mean

exceeds it only on CIFAR-10 at 1000:1 and ϵ=1 (0.0029 vs. 0.0027, with 10 test anomalies per draw).

Table 3: Radius-only release (standard instance-level ϵ-DP for the whole record) across imbalance ratios: mean AUPRC / bound $e^{\epsilon}/(e^{\epsilon}+r)$, on pools drawn from the test split (Credit Card: three encoder seeds $\times$ 20 draws; CIFAR-10: deployed encoder $\times$ 20 draws).

Ratio	ϵ=1	ϵ=5	ϵ=10	no noise
Credit Card				
10:1	0.176 / 0.214	0.757 / 0.937	0.879 / 1.000	0.935
50:1	0.041 / 0.052	0.522 / 0.748	0.834 / 0.998	0.910
100:1	0.021 / 0.026	0.382 / 0.597	0.809 / 0.995	0.899
500:1	0.004 / 0.005	0.132 / 0.229	0.735 / 0.978	0.868
577:1	0.004 / 0.005	0.123 / 0.204	0.723 / 0.974	0.863
CIFAR-10				
10:1	0.142 / 0.214	0.501 / 0.937	0.630 / 1.000	0.794
50:1	0.032 / 0.052	0.260 / 0.748	0.464 / 0.998	0.634
100:1	0.016 / 0.026	0.165 / 0.597	0.369 / 0.995	0.559
500:1	0.005 / 0.005	0.056 / 0.229	0.184 / 0.978	0.371
1000:1	0.003 / 0.003	0.039 / 0.129	0.150 / 0.957	0.339

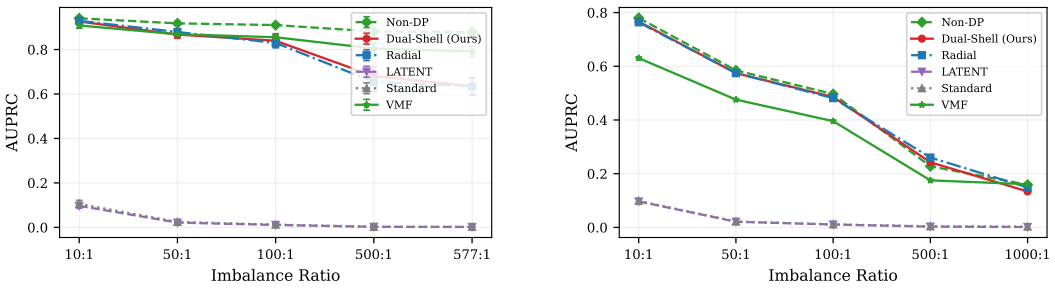
Scoped release, both metrics. Figure 3 shows the scoped per-record release (direction public) on both metrics: Norm-AUPRC (anomaly score $= \|z_{priv}\|$) and LR-AUPRC (a logistic-regression classifier trained on the released representations). LATENT stays near-random on both metrics (0.010 Norm-AUPRC at ϵ=1 on CIFAR-10); Standard Isotropic recovers LR-AUPRC only at relaxed budgets; VMF's high Norm-AUPRC reflects the unprotected norm. Over three encoder seeds (five noise repetitions each), the logistic-regression AUPRC at $\epsilon \in \{5, 10, 20\}$ is $0.81/0.85/0.86$ for ORBIT, $0.80/0.85/0.86$ for the radial ablation, and $0.008/0.08/0.59$ for Standard on Credit Card; on Covertype it is $0.72/0.76/0.75$, $0.72/0.77/0.77$, and $0.008/0.08/0.47$ (non-private: 0.86 and 0.77). Releasing the direction alone gives LR-AUPRC 0.79 on Credit Card and 0.69 on Covertype (three seeds). A classifier given the direction and the released radius as separate features stays at this level. One trained on z_{priv} is lower at $\epsilon \leq 2$ (0.64 and 0.61 at ϵ=1), because the noisy radius rescales the direction features, and higher from ϵ=5 on. At $\epsilon \geq 50$ the radial ablation slightly outperforms ORBIT on CIFAR-10 in Norm-AUPRC (a *gravity tax*: compressing extreme outliers slightly reorders rankings when noise is negligible).



(a) Credit Card (d=32) (b) CIFAR-10 (d=128)

Figure 3: Scoped per-record release (direction public) across five DP mechanisms, depicting Norm-AUPRC (left panels) and LR-AUPRC (right panels). Standard uses the per-coordinate calibration ($\sqrt{d}\,\epsilon$-DP for the vector; a favorable reference).

Imbalance sensitivity in the scoped release (RQ4). Figure 4 reports the full imbalance sweep (ϵ=1, 20 repetitions per ratio; the classifier trains on a pool drawn from the training split and is evaluated on a pool drawn from the test split, both at the given ratio). The two mechanisms that protect the radial magnitude are within one standard deviation of each other at every ratio on both datasets. Both reach high AUPRC through the public direction (0.64 at 577:1 on Credit Card), as VMF does through its unprotected norm, whereas LATENT and Standard Isotropic fall to the base rate as imbalance grows ($\leq$ 0.11 throughout).

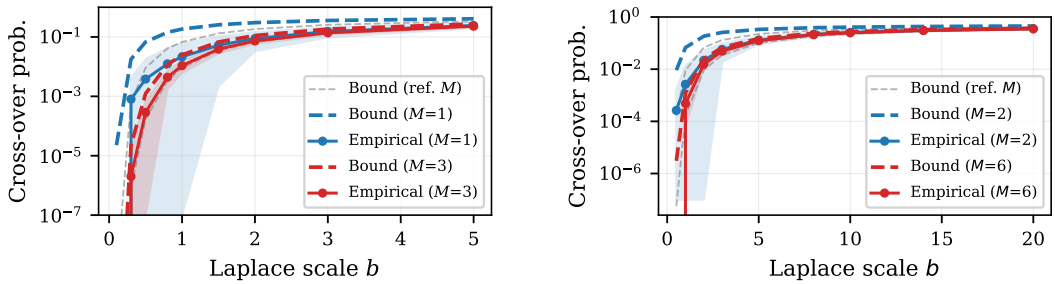


(a) Credit Card (d=32)

(b) CIFAR-10 (d=128)

Figure 4: Imbalance sensitivity in the scoped release (RQ4, $\epsilon = 1$, 20 repetitions; evaluation pools from the test split) across five DP mechanisms. The two radial mechanisms are within one standard deviation at every ratio and draw their utility from the public direction; LATENT and Standard (per-coordinate calibration) stay near-random; VMF's high AUPRC reflects its unprotected norm.

Crossover bound (RQ2). Figure 5 shows the empirical crossover probability against the bound of Theorem 3 (Section 6.3).



(a) Credit Card (d=32): $M \in \{1, 3\}$ focal, $M \in \{2, 4\}$ reference.

(b) CIFAR-10 (d=128): $M \in \{2, 6\}$ focal, $M \in \{4, 8\}$ reference.

Figure 5: Theorem 3 validation (RQ2): empirical crossover probability (mean $\pm 1\sigma$ over up to 200 anomaly records that satisfy the premise $\|z\|_2 > R_2$, 5,000 Laplace draws each) against the bound $\frac{1}{2}\exp(-M/b)$ (dashed).

Margin at fixed sensitivity (RQ3). Figure 6 shows the margin sweep of Section 6.3. For each $M \in \{0, 0.5, 1, 2, 3, 4\}$ the Credit Card encoder is retrained with the squared shell-targeting loss of Equation (2) (R_1=2, R_2=2+M, R_{max}=8 fixed, $\lambda = 5$, stable recipe, seeds 42/43/44), and every training and test record is released with the two-sided projection and $b = 7/\epsilon$. Crossover is measured on the test anomalies as the fraction of pre-clipping noisy radii r_{priv} below R_1. The AUPRC of the released radius rises with M at every ϵ (left panel). The LR-AUPRC of a classifier on the full vector does not (right panel): the public direction sets it at every M (0.54–0.85).

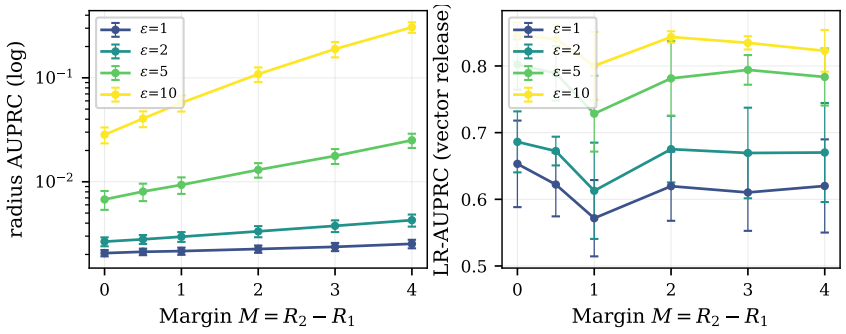
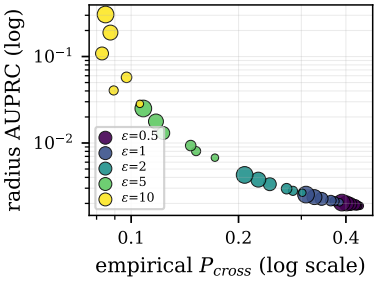
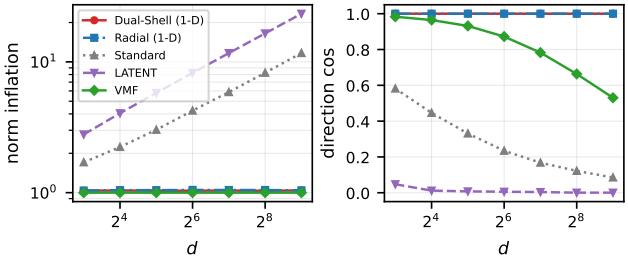


Figure 6: Margin sweep at fixed sensitivity (Credit Card, R_{max}=8, Δh=7 for every M; mean $\pm$ std over three training seeds $\times$ three noise repetitions). The AUPRC of the released radius (left, log scale) increases with M at every ϵ, with no interior optimum over $M \in [0, 4]$; the LR-AUPRC of the vector release (right) is set mainly by the public direction.

Crossover–utility relation and distortion across dimensions. Figure 7 shows the 30 configurations of the margin sweep against the measured crossover probability (a) and the distortion of synthetic shell vectors across latent dimensions at ϵ=10 (b).



(a) Margin sweep, radius AUPRC against crossover (log–log): 30 (M, ϵ) configurations, mean over three training seeds (color $= \epsilon$, size $= M$); $r = -0.94$.



(b) Dimension scaling (ϵ=10): only radial (1-D) noise is constant in d.

Figure 7: Mechanistic validation (Credit Card) and dimension scaling.

Dimension and utility. Figure 8 shows the per-record AUPRC of encoders retrained at each latent dimension with the same geometry (R_1=2, R_2=5, R_{max}=8, μ_{max}=0.5; three training seeds, three noise repetitions). Standard Isotropic is near-random at $\epsilon \leq 5$ for every d; the radial mechanism's Norm-AUPRC is constant in d (0.017–0.018 at ϵ=5, 0.18–0.19 at ϵ=10, 0.75–0.76 at ϵ=20). The classifier AUPRC of the radial mechanisms does vary with d (ϵ=5: 0.37 at d=8, 0.79 at d=32); because the radial noise is identical across d, this variation comes from the representation learned at each d, not from the noise scale.

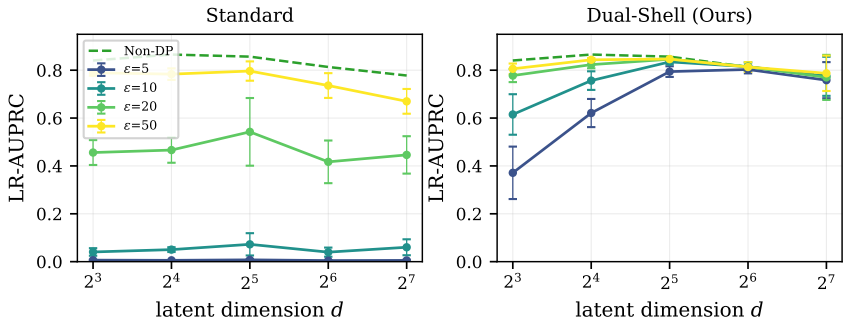


Figure 8: Per-record LR-AUPRC against the latent dimension (Credit Card, encoders retrained at each d; mean $\pm$ std over three training seeds). Standard Isotropic is near-random at $\epsilon \leq 5$ for every d.

Fully synthetic release, full table. Table 4 lists the values summarized in Section 6.2.

Table 4: Fully synthetic release: TSTR AUPRC (mean $\pm$ std over three encoder seeds $\times$ ten repetitions). *Latent*: classifier trained on synthetic latents and tested on the encoded real test split (DP by post-processing). *Decoded*: synthetic latents decoded by the fixed, non-private curator decoder (decodability only); the decoder does not reproduce the Covertype minority even without privacy (Non-DP 0.10), so no decoded Covertype values are shown. †DP generators produce raw feature rows (ϵ covers generator training); their decoded column evaluates those rows directly, and their latent column encodes them with the curator encoder. ‡PATE-GAN (synthcity defaults) reaches its iteration cap before spending ϵ=1, so its ϵ=1 and ϵ=10 runs coincide (realized $\epsilon \leq 1$); three repetitions.

	Latent		Decoded	
Method	**ϵ=1**	**ϵ=10**	**ϵ=1**	**ϵ=10**
Credit Card (577:1)				
Non-DP (shaped encoder)	0.859 ± 0.022		0.633 ± 0.071	
ORBIT	0.862 ± 0.012	0.869 ± 0.009	0.567 ± 0.095	0.639 ± 0.066
Radial (no gravity)	0.860 ± 0.012	0.866 ± 0.011	0.580 ± 0.071	0.647 ± 0.049
Standard Isotropic	0.838 ± 0.035	0.861 ± 0.015	0.183 ± 0.203	0.676 ± 0.028
DP-MoG†	0.717 ± 0.066	0.709 ± 0.037	0.503 ± 0.089	0.683 ± 0.022
PATE-GAN†‡, full table	0.110 ± 0.107		0.190 ± 0.176	
Non-DP, unshaped encoder	0.647 ± 0.052		0.681 ± 0.033	
Covertype (500:1)				
Non-DP (shaped encoder)	0.758 ± 0.057		–	
ORBIT	0.770 ± 0.038	0.737 ± 0.070	–	–
Radial (no gravity)	0.776 ± 0.033	0.733 ± 0.073	–	–
Standard Isotropic	0.780 ± 0.027	0.799 ± 0.024	–	–
DP-MoG†	0.223 ± 0.226	0.260 ± 0.255	–	–
PATE-GAN†‡, full table	0.134 ± 0.140		–	
Non-DP, unshaped encoder	0.169 ± 0.078		–	

Mechanism isolation in the synthesis path. Table 5 separates three candidate explanations for the synthetic-release results on Credit Card. *Direct* trains the classifier on the released training latents (or their decoded rows) without a density model. *Std-proj* rescales Standard's released latents to $\mathrm{clip}(\|z\|_2, R_1, R_{max})$ before density estimation (post-processing). VMF keeps the norm and perturbs the direction. Direct and density-based training give similar latent-space AUPRC, so density estimation does not explain the retained utility. Restoring the norm does not restore Standard's decodability at ϵ=1. F1 at the default threshold does not order the mechanisms consistently: Standard has the lowest F1 at ϵ=1 and the highest at ϵ=10, above the non-private value, so F1 here reflects where the threshold falls on the real test distribution rather than preserved structure. On Covertype, all latent-space values lie within 0.75–0.80 and the decoder does not reproduce the minority (Non-DP 0.10).

Table 5: Mechanism isolation in the synthesis path (Credit Card; mean over three encoder seeds × ten repetitions; ϵ=1 / ϵ=10).

Variant	Latent AUPRC	Decoded AUPRC	Latent F1
Non-DP	0.859	0.633	0.351
ORBIT, density model	0.862 / 0.869	0.567 / 0.639	0.292 / 0.346
ORBIT, direct	0.871 / 0.867	0.698 / 0.713	0.303 / 0.188
Standard, density model	0.838 / 0.861	0.183 / 0.676	0.091 / 0.851
Standard, direct	0.846 / 0.861	0.176 / 0.685	0.080 / 0.854
Std-proj, density model	0.840 / 0.860	0.151 / 0.642	0.189 / 0.791
VMF, density model	0.860 / 0.865	0.530 / 0.569	0.168 / 0.614

CIFAR-10: F1 under latent-space TSTR and L_2-norm inflation. Figure 9 shows the latent-space TSTR curves on CIFAR-10 over three encoder seeds. AUPRC stays close to the non-private value for every mechanism. F1 at the default threshold is 0.00 for Standard at ϵ=1 (0.16 for ORBIT, 0.19 without privacy) but 0.54–0.55 for Standard at $\epsilon \geq 10$ (0.12–0.14 for ORBIT), so, as on Credit Card, F1 does not order the mechanisms consistently. Figure 10 visualizes the failure mechanism through the (fixed, curator-side) decoder: injecting independent Laplace noise into 128 dimensions inflates the latent L_2-norm from 15.4 to $\approx$ 35.4 at ϵ=20 (and to $\approx$ 65.3 at ϵ=10) and randomizes the direction, producing shattered, posterized color blocks; on Credit Card, restoring the norm alone does not restore decodability (Section 6.2), so the artifacts cannot be attributed to norm inflation alone, whereas radial perturbation keeps the norm bounded and the decoder reconstructs natural anomaly silhouettes. The baseline blurriness of the Non-DP reconstructions likely reflects the weight on the margin penalty ($\lambda = 50$) relative to reconstruction in Stage 1. These grids are qualitative evidence of geometric integrity only; as stated in Section 6.2, pixel-space TSTR collapses to the base rate for every method including Non-DP, so no pixel-level generative claim is made.

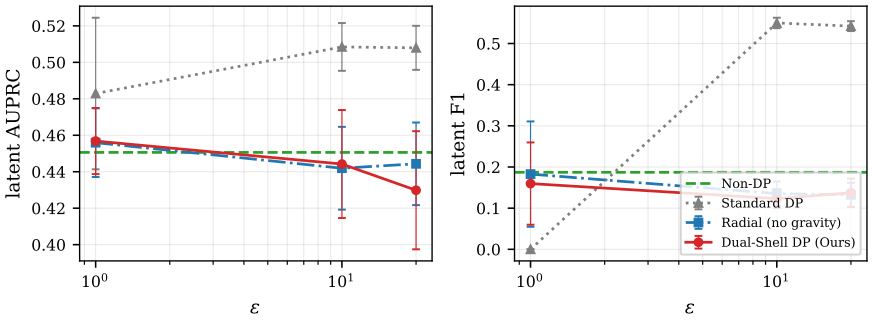
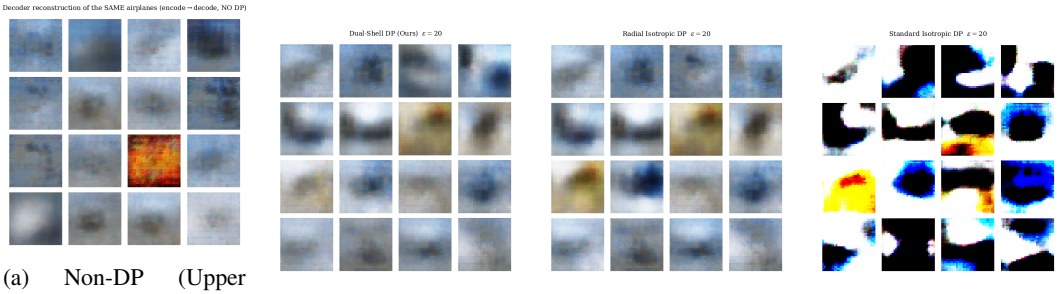


Figure 9: Latent-space TSTR on CIFAR-10 (mean ± std over three encoder seeds × ten repetitions). AUPRC stays close to the non-private value for every mechanism; F1 at the default threshold changes order between ϵ=1 and $\epsilon \geq 10$.

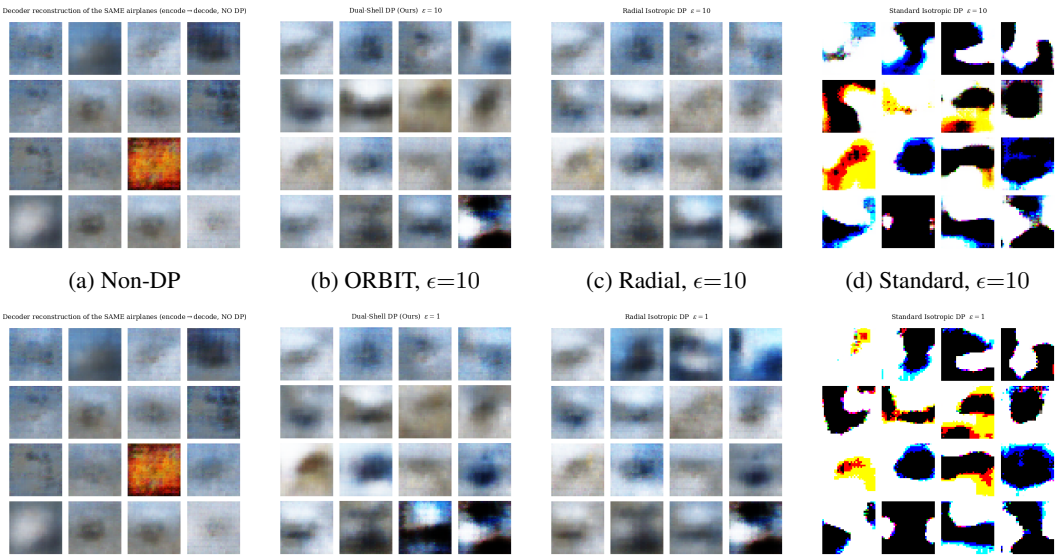


(a) Non-DP (Upper Bound) (b) ORBIT (Ours) (c) Radial (d) Standard Isotropic

Figure 10: Qualitative TSTR reconstructions ($\epsilon = 20$). Standard Isotropic (d) inflates the L_2 norm and randomizes the direction, producing saturated artifacts. ORBIT (b) perturbs only the norm and preserves the airplane silhouettes with fidelity comparable to the non-private reconstructions (a). Grids at tighter budgets ($\epsilon = 1$ and $\epsilon = 10$) are in Appendix E.

E Supplementary Qualitative CIFAR-10 Grids at Tighter Budgets

Figure 11 complements the ϵ=20 grids of Figure 10 (Appendix D) with decoded samples at ϵ=10 and ϵ=1, all through the same fixed curator-side decoder. At ϵ=10, Standard Isotropic does not recover the anomaly manifold: independent perturbations across 128 dimensions push the latents far outside the decoder's training range (norm $\approx$ 65.3) and randomize their direction, whereas the radial mechanisms keep the direction and the decoder produces recognizable silhouettes. At ϵ=1, ORBIT's noise scale ($b = 16$) is close to the target anomaly norm (≈ 17), so the radial mechanisms produce blurrier samples, but the clip of the released radius keeps them in the decoder's range and the direction unchanged; Standard Isotropic degrades into saturated artifacts. These grids are qualitative evidence of geometric integrity only, consistent with the per-record collapse at $\epsilon \leq 1$ in Figure 1.



(a) Non-DP (b) ORBIT, ϵ=10 (c) Radial, ϵ=10 (d) Standard, ϵ=10

(e) Non-DP (f) ORBIT, ϵ=1 (g) Radial, ϵ=1 (h) Standard, ϵ=1

Figure 11: Qualitative decoded grids on CIFAR-10 at ϵ=10 (top) and ϵ=1 (bottom). Radial mechanisms keep latents on the decoder manifold at both budgets; Standard Isotropic does not.

F Semi-Synthetic Protocol (Robustness)

The main text evaluates the fully synthetic release, in which no raw record enters downstream training. A complementary semi-synthetic protocol combines 2,000 synthetic minority rows with a

20,000-row subsample of *real, unprotected* majority rows. We run it with the pipeline of Section 6.2 (same releases, density models, and classifier; only the minority passes through the mechanism). Table 6 reports the results. At ϵ=1 the radial mechanisms keep most of their latent-space utility under this protocol (ORBIT 0.84 and radial ablation 0.84 vs. 0.85 without privacy on Credit Card; 0.68 and 0.68 vs. 0.75 on Covertype), as does Standard (0.84 and 0.79). Without the clip of the released radius (Appendix A), the inverted directions reduce ORBIT to 0.35 on Credit Card and 0.31 on Covertype. In decoded space at ϵ=1, ORBIT (0.62) is above Standard (0.41) and DP-MoG (0.59) on Credit Card. PATE-GAN trained on the minority partition reaches 0.69 in latent space and 0.55 on its raw rows at ϵ=1 on Credit Card. At ϵ=10 all mechanisms are close to the non-private values. This protocol is a robustness check; the main evidence is the fully synthetic release.

Table 6: Semi-synthetic protocol (synthetic minority + real majority rows): TSTR AUPRC, mean over three encoder seeds $\times$ ten repetitions (ϵ=1 / ϵ=10). Decoded Covertype values are limited by the decoder (Non-DP 0.28 with real majority rows). †Trained on the minority partition; its decoded column evaluates its raw rows; three repetitions.

	Credit Card		**Covertype**	
Method	Latent	Decoded	Latent	Decoded
Non-DP	0.847	0.677	0.751	0.282
ORBIT	0.842 / 0.862	0.624 / 0.671	0.681 / 0.733	0.242 / 0.275
Radial (no gravity)	0.841 / 0.860	0.626 / 0.672	0.676 / 0.733	0.234 / 0.270
Standard Isotropic	0.842 / 0.861	0.407 / 0.551	0.785 / 0.799	0.012 / 0.020
DP-MoG	0.764 / 0.759	0.591 / 0.707	0.597 / 0.678	0.187 / 0.252
PATE-GAN†	0.693 / 0.765	0.547 / 0.634	0.621 / 0.695	0.296 / 0.286

G Sensitivity Projection Analysis

Theorem 1 relies on the deterministic feasible-domain projection of Equation (3), which enforces both radial bounds before noise injection. A natural concern is whether this projection changes the method's empirical behavior. We therefore measure its effect on the evaluated Credit Card (three stable-recipe encoders) and CIFAR-10 representations; the CIFAR-10 setup is detailed in Section 6 and Appendix C. We report three quantities. The lower-clip rate is the fraction of records with $\|z\|_2 < R_1$. The h-modified rate is the fraction of records whose protected quantity changes under the projection, that is, $h(\hat{z}) \neq h(z)$. On the lower side, this happens exactly when $\|z\|_2 < R_1 - \mu_{max}$, because for $R_1 - \mu_{max} \leq \|z\|_2 < R_1$ the gravity term already maps the norm to R_1. The upper-clip rate is the fraction with $\|z\|_2 > R_{max}$.

Table 7 summarizes the results. On the three Credit Card encoders, the projection changes the protected quantity of 1, 4, and 0 of 284,807 records. The upper clip accounts for 1, 3, and 0 of these (largest norms 8.20, 8.29, and 7.78 for R_{max}=8); one record on the second encoder falls below $R_1 - \mu_{max}$. No CIFAR-10 record is changed, in either the training or the test split, and the largest CIFAR-10 norm is 19.36 (R_{max}=20). No record in either dataset has $z = 0$. We also ran a paired comparison of downstream utility with and without the projection, using identical Laplace noise (Credit Card at $\epsilon \in \{1, 10\}$, CIFAR-10 at $\epsilon \in \{10, 20\}$). On Credit Card, Norm-AUPRC changes by at most 0.004 and LR-AUPRC by at most 0.002, and both are identical on the encoder where no record is changed. On CIFAR-10, Norm-AUPRC changes by at most 0.003 and LR-AUPRC by at most 0.001. The protected quantity differs by less than 10^{-14} between the two runs, so these residuals come from ties among records whose released radius is clipped to R_1, which rounding orders differently, and from the classifier fit. On the evaluated representations, the projection therefore has a small empirical effect while enforcing the bounded radial domain required by Theorem 1.

Table 7: Feasible-domain projection analysis on the evaluated representations (Credit Card: stable-recipe encoder, seed 42; the paired $|\Delta|$ is the maximum over three encoders). The projection changes the protected quantity for at most four Credit Card records per encoder and for no CIFAR-10 record, and no record has $z = 0$.

Dataset	**Lower clip** ($\|z\| < R_1$)	**h-modified** ($\|z\| < R_1 - \mu_{max}$)	**Upper clip** ($\|z\| > R_{max}$)	$z = 0$ (count)	$\|\Delta\|$**AUPRC** (paired)
Credit Card	50.53%	0.0000%	0.0004%	0	≤ 0.004
CIFAR-10 train	46.09%	0.0000%	0.0000%	0	≤ 0.003
CIFAR-10 test	46.52%	0.0000%	0.0000%	0	

H Empirical Validation of Directional Concentration

We now check the concentration part of Assumption 1, the premise of the directional bound (Section 5); whether identities within a class share one law is examined by the audit of Section 6.4. Assumption 1 posits that after Stage-1 convergence, the per-class direction distribution is well approximated by a shared von Mises–Fisher (vMF) law with a large concentration parameter κ_c.

Per-class concentration. Table 8 reports, for each class, the mean resultant length $\bar{R} = \|\frac{1}{n}\sum_i \vec{u}_i\|$ and the corresponding vMF concentration estimate $\hat{\kappa} = \bar{R}(d - \bar{R}^2)/(1 - \bar{R}^2)$. All classes are sharply concentrated: directions deviate from their class-mean by a median angle of only $5.3°$–$16.4°$, far below the $\approx 90°$ expected for isotropic directions in high dimensions. The effect is most pronounced in the high-dimensional CIFAR-10 latent space ($d = 128$), where the normal class attains $\hat{\kappa} \approx 8.7 \times 10^3$. The angle-to-mean histograms (Figures 13a and 13b) confirm a tight unimodal clustering consistent with a vMF law.

Concentration during training. Figure 12 tracks the vMF concentration $\hat{\kappa} = \bar{R}(d-\bar{R}^2)/(1-\bar{R}^2)$ (log scale) across Stage-1 epochs. On Credit Card, the concentration of the well-populated normal class rises steadily ($\hat{\kappa} : 1.7 \times 10^2 \to 2.8 \times 10^2$, deployed encoder, checkpoints every five epochs). On CIFAR-10 (deployed run, checkpoints every five epochs), $\hat{\kappa}$ is highest in the first epochs, when most images map to nearly the same direction, and decreases as the representation forms, settling at $\hat{\kappa} \approx 8.7 \times 10^3$ (normal) and $\approx 2.0 \times 10^3$ (anomaly). (Because $\hat{\kappa}$ is numerically sensitive as $\bar{R} \to 1$ in high dimensions, exact magnitudes vary across runs; Table 8 reports the per-class values for the deployed encoder.) Assumption 1 concerns the converged encoder, and representations on both datasets exhibit sharp angular concentration at convergence. Concentration alone does not rule out leakage: on CIFAR-10, the anomaly class is sharply concentrated, yet the direction-only audit of Section 6.4 still detects membership, which is why Lemma 1 is paired with the audit rather than treated as a standalone guarantee.

Table 8: Angular concentration of latent directions (deployed encoder, full data). High $\hat{\kappa}$ and small angles to the mean support the concentration premise of Assumption 1; they do not by themselves rule out identity leakage (Section 6.4). Note that $\hat{\kappa}$ is a plug-in estimate that becomes sensitive as $\bar{R} \to 1$ in high dimension; the mean resultant length $\bar{R}$ and the median angle are the robust concentration indicators.

Dataset	**Class**	n	$\bar{R}$	$\hat{\kappa}$	**median** $\angle$
Credit Card	normal	284,315	0.947	283	$16.1°$
Credit Card	anomaly	492	0.841	90	$16.4°$
CIFAR-10	normal	54,000	0.993	8,741	$5.3°$
CIFAR-10	anomaly	540	0.968	1,953	$8.8°$

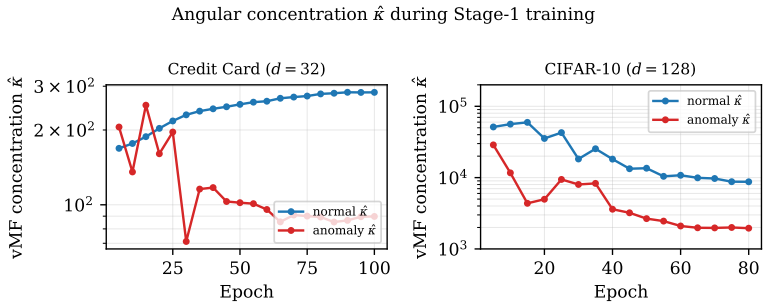
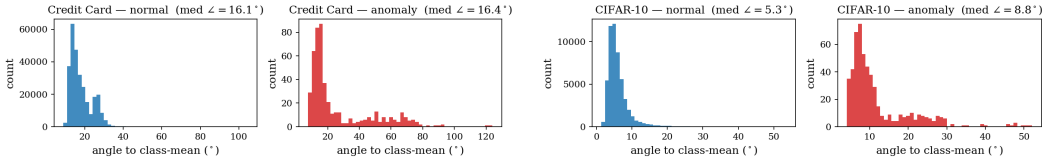


Figure 12: vMF concentration $\hat{\kappa}$ (log scale) of latent directions during Stage-1 training. On Credit Card, the normal-class concentration grows during training; on CIFAR-10 (checkpoints every five epochs) it is highest during the initial epochs and settles at a high value as the representation forms. (As $\hat{\kappa}$ is numerically sensitive when $\bar{R} \to 1$ in high d, exact magnitudes vary across runs; cf. Table 8 for the stable deployed-encoder values.)



(a) Credit Card (b) CIFAR-10

Figure 13: Distribution of the angle between each latent direction and its class-mean direction. The tight unimodal clustering near $0°$ is consistent with a highly concentrated vMF distribution (Assumption 1).

Functional roles of the two coordinates. The direction and the radius play different functional roles but are not disentangled. Table 9 measures the class information each coordinate carries on the test split: a radius-only score (AUC and AUPRC of $\|z\|_2$) and a direction-only probe (logistic regression on $\vec{u}$, balanced classes, five seeds), each for the shaped encoder and for an unshaped encoder with the same architecture and recipe, averaged over three encoder seeds. Both coordinates separate the classes well in AUC. The radius does so by construction, because $\mathcal{L}_{margin}$ pushes the two classes toward different shells; shaping raises its AUC most on CIFAR-10 ($0.60 \to 0.91$). AUC does not reflect rare-class precision, however: the radius-only AUPRC rises from 0.02–0.15 without shaping to 0.54–0.86 with it on all three datasets, so shaping is what places the rare-class separation on the radius. Shaping also keeps the norms inside the design bound: without it, 61–73% of tabular records exceed R_{max}=8, so the projection would change the protected quantity of most records. On the tabular datasets the direction separates the classes about as well without shaping, so there its class information does not come from shaping; on CIFAR-10 shaping raises it ($0.83 \to 0.92$). Shaping concentrates within-class directions by a factor of about 1.5 on the tabular datasets and about 17 on CIFAR-10 (median angle of normal-class directions to their class mean). Beyond discrete class identity, the radius retains each record's position within its shell. Within the anomaly class, this position is positively correlated with two separately computed atypicality measures: an isolation-forest score fitted on raw normal records (Spearman $\rho = 0.33$–0.43 on Credit Card and 0.32–0.37 on Covertype across the three encoder seeds) and the standardized distance from the normal mean ($\rho = 0.36$–0.45 and 0.08–0.12). On Credit Card it is not significantly correlated with the transaction amount ($|\rho| < 0.06$, $p > 0.2$). Within the normal class, the radius is nearly constant (standard deviation 0.03 on Credit Card and 0.06 on Covertype). The protected quantity therefore includes the public class margin and, within the anomaly class, a radial position that is positively correlated with atypicality; the latter is the quantity we designate as privacy-sensitive. In a multiclass setting, the direction could analogously indicate the class and the radius a graded property such as confidence or typicality. Even there the two components need not be independent, since typicality can differ

across classes, so we describe them by the roles they play in the release rather than as disentangled semantics.

Table 9: Class information in each coordinate (test split; shaped / unshaped encoder; mean over three encoder seeds). Direction probe: logistic regression on $\vec{u}$, mean over five probe seeds. Angle: median angle of normal-class directions to their class mean (full data).

Dataset	Radius AUC	Radius AUPRC	Direction AUC	Normal-class angle
Credit Card	0.962 / 0.953	0.863 / 0.150	0.976 / 0.972	16.5° / 27.2°
Covertype	0.995 / 0.920	0.825 / 0.018	0.997 / 0.990	29.6° / 43.7°
CIFAR-10	0.907 / 0.596	0.542 / 0.021	0.922 / 0.830	4.8° / 83.1°

I Radial Norm Density under Perturbation

Figure 14 shows the released radius r_{priv} of the radius-only release on Credit Card (deployed encoder; 20 noise draws per record) for all test anomalies (top) and for 20,000 random normal test records (bottom). The crossover that limits precision is normal records released above R_2. At ϵ=5 this is 5.9% with the default window, 7.7% for the gravity-free radial release, and 0.35% with the window $[R_1, R_2]$; at ϵ=10 it is 0.7%, 1.2%, and below 0.01%. With 580 normal records per anomaly in the test split, precision is governed mainly by this fraction, consistent with the gap between the two windows in Figure 1 (0.02 vs. 0.12 AUPRC at ϵ=5). The window $[R_1, R_2]$ also releases fewer anomalies below R_1 (6.1% vs. 11.1% at ϵ=5); 5.1% of the test anomalies already lie at the inner shell without privacy. At ϵ=1 the noise spreads both classes across the margin for every window.

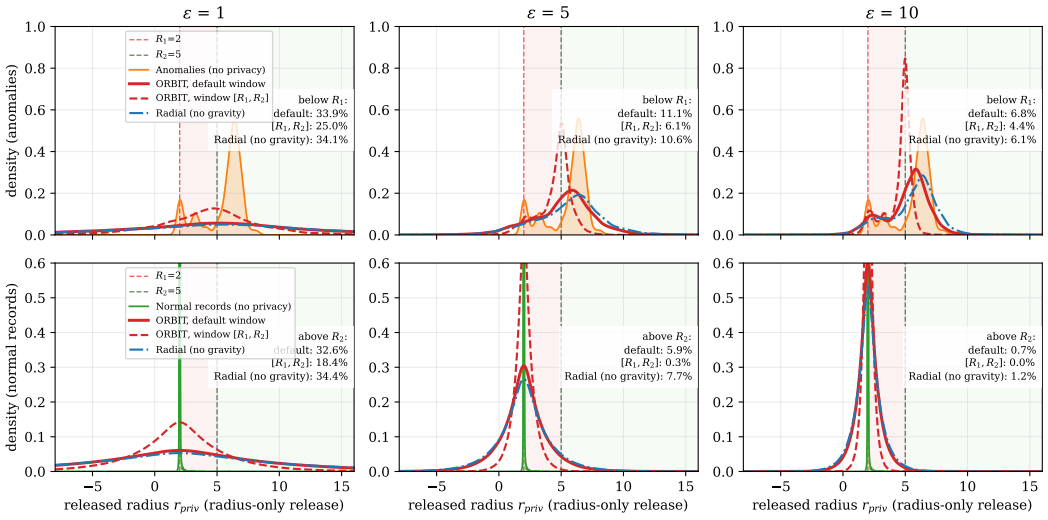


Figure 14: Released radius r_{priv} of the radius-only release on Credit Card (deployed encoder) for test anomalies (top) and normal test records (bottom), with the fraction released across the margin. The window $[R_1, R_2]$ (Δh=M) releases far fewer normal records above R_2 than the default window or the gravity-free radial release.